Universiteit Utrecht

Bachelor thesis

Mathematisch Instituut

Markov Decision Processes and its Applications in Medical Sciences

Author
Peggy Bergman
5834783

Supervisor
Dr. K. Dajani

2nd January 2020

Contents

Acknowledgements

I would like to thank my supervisor Dr. Karma Dajani for suggesting Markov decision processes as a subject for my thesis. Further, I would like to give my gratitude for her guidance during the writing process.

Introduction

Markov decision processes are important stochastic processes, since it is divers in its applications and Markov decision processes are very useful for studying optimization problems. Even in disciplines outside mathematics, such as economics, computer sciences and many more, Markov decision processes are used. In this thesis, we will apply Markov decision processes in the medical sciences. We will show that it has its applications for modeling the spread of infectious disease and for the treatment of ischemic heart disease.

In this thesis, we will discuss the theory behind Markov decision processes, more specifically discrete-time Markov decision processes. This will be done in steps, starting with chapter 1 which discusses the sequential decision model. This chapter is based on chapter 1 of Puterman ([6]), with the exception of the examples. The purpose of this chapter is to learn about the basics behind a decision process. In chapter 2 we will explain the theory of Markov chains, since Markov decision processes are built upon Markov chains. Chapter 2 is based on chapter 4 of Ross ([7]) and Appendix A of Puterman ([6]), with the exception of the examples and proofs of propositions (2.3.1) and (2.3.2) and corollary (2.2.2). Proposition (2.3.2) is based on pages 96-97 of Kulkarni ([5]). Chapter 3 gives an outline of Markov decision processes, here we will discuss the formulation of the model, finite-horizon and infinite-horizon problems. This chapter is based on Chapter 2, 3, 4 and 5 of Puterman ([6]) and on Chapter 5 and 6 of Taylor ([9]). In chapter 4 we will discuss partially observed Markov decision processes, this chapter is based on Chapter 7 of Krishnamurthy ([4]), Hauskrecht ([3]) and Cassandra ([1]), with the exception of example (4.0.1). This is roughly the outline for the first four chapters. These four chapters will give the mathematical background needed for chapter 5. In chapter 5, we will elaborate on the applications in medical sciences. Chapter 5 is based on the articles of Yaesoubi and Cohen ([10]) and ([11]) and on the article of Hauskrecht ([3]), with the exception of the example.

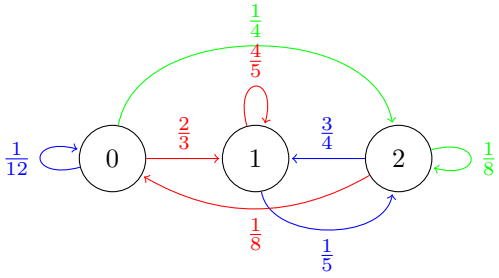
Chapter 1

The Sequential Decision Model

In this chapter, we will explain the sequential decision model and we will discuss how this model plays a part in the following chapters.
The sequential decision model is described as follows. A decision maker observes the state of a system at a specified point in time. Based on the observed state, he chooses an action. This chosen action produces two results, namely the decision maker receives an immediate reward and the system evolves to a new state at a subsequent point in time according to a probability distribution, which is determined by the choice of action. Arrived at this new point in time, the decision maker faces a similar problem, however the system may be in a different state and it may be possible for him to choose from a different set of actions. So the key elements of the sequential decision model are the following:

1. A set of decision times;
2. A set of system states;
3. A set of available actions;
4. A set of state and action dependent rewards;
5. A set of state and action dependent transition probabilities.

Example 1.0.1. Assume for a moment a frog in a pond with three lilies. This frog is the decision maker, he will decide what actions to take. The lilies in the pond will represent the different states and the food by the lilies the reward. The jump from one lily to the other will represent the action and the probability of this jump is the transition probability.



The rewards are +1 in state 0, +5 in state 1 and +2 in state 2, the rewards will represent the amount of food. Further, we assume that time runs from 0 to 10 minutes. Therefore, our set of decision epochs is $\{0,1,2,...,10\}$, the set of states is $\{0,1,2\}$, the set of immediate rewards is $\{1,2,5\}$, the set of actions is $\{R,B,G\}$, with R for the red line, B for the blue line and G for the

green line. So the action is whether the frog jumps to another lily or not. And lastly, the set of state transition probabilities are given by $\{P_{ij}\}$, which gives us the probability that the frog jumps from state i to state j with $i, j \in \{0, 1, 2\}$ according to the diagram. If there is no arrow between state i and state j, then the probability of that action is zero.

♦

Given a sequential decision model, we would like to know how the decision maker can maximize his outcome. So we will be interested in finding the optimal policies and decision rules. A **decision rule** specifies the action to be chosen at a particular time. It may depend on the present state alone or together with all previous states and actions. And a **policy** provides the decision maker with a prescription for choosing an action in any possible future state. So a policy is actually a sequence of decision rules.

In the following chapters, we will discuss two particular sequential decision models, namely the Markov Decision Processes (MDPs) and Partially Observed Markov Decision Processes (POMDPs). In these models, the set of available actions, the set of rewards and the set of transition probabilities depend only on the current state and action and not on the states occupied and action chosen in the past. So in other words, the state transitions of a Markov Decision Process and a Partially Observed Markov Decision Process all satisfy the Markov property. This means that they satisfy the following condition: Let X_n denote the state in time period n, then

$$\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})$$

for $n \in \mathbb{N}$ point in time.

Further, for MDPs we will make the assumption that all of the key elements of the sequential decision model are known to the decision maker at the time of each decision. When we move on to POMDPs this assumption will no longer hold. Thus, in a MDP the decision maker has all the information that is available. Hence, the rewards and the states are perfectly observed. In a POMDP, the decision maker is unsure in which state he is. He only receives an observation and the reward. But before we start with the more advanced models, we will take a look at the simplest Markov model, namely Markov chains.

Chapter 2

Markov Chains

In this chapter, we will discuss some basic theory of discrete-time Markov chains that is relevant for the analysis of Markov Decision Processes (MDPs) and Partially Observed Markov Decision Processes (POMDPs).

2.1 Introduction to Markov Chains

In this section, we will introduce Markov chains. We start with the assumption that we have a process that takes a value in each given time period. Let X_n denote the value in time period n. Also, suppose that we want to make a probability model for the sequence of the successive values $X_0, X_1, ...$ Now, we can make the assumption that the conditional distribution of X_{n+1} given $X_n, X_{n-1}, ..., X_0$ depends only on X_n. So the future state depends only on the present state and not on past states. Such an assumption is called memoryless and a stochastic process that satisfies this assumption is defined as a **Markov chain**. We have seen this assumption before in chapter 1 as the Markov property. Therefore, the Markov chain satisfies the Markov property

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, ..., X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

for $n \in \mathbb{N}$ point in time. Further, we call a Markov chain **time-homogeneous** if

$$\mathbb{P}(X_{n+1} = j \mid X_n = i)$$

does not depend on n. From now on we assume that the Markov chain is time-homogeneous. So formally, we have a stochastic process $\{X_n \mid n = 0, 1, 2, ...\}$ that takes on a finite or countable number of possible values. We assume that whenever the stochastic process is in state i, $X_n = i$, there is a probability P_{ij} that says that the process will be in state j next, so $X_{n+1} = j$. The probability P_{ij} is called the **one-step transition probability**. Further, we assume that this probability is defined as

$$P_{ij} \coloneqq \mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

for all states $i_0, i_1, ...i_{n-1}, i$ and j and for all $n \geq 0$, so the stochastic process is time-homogeneous. Therefore, we can rewrite P_{ij} as

$$P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_1 = j \mid X_0 = i).$$

Because P_{ij} is a conditional probability, we know that $P_{ij} \geq 0$ and

$$\sum_{j=0}^{\infty} P_{ij} = \sum_{j=0}^{\infty} \mathbb{P}(X_1 = j \mid X_0 = i) = 1,$$

since j is running over all the possible values of X_1 for $i = 0, 1, 2, ...$ The **transition matrix P** of an one-step transition probability is defined as:

$$\mathbf{P} \coloneqq \begin{pmatrix} P_{00} & P_{01} & P_{02} & \dots \\ P_{10} & P_{11} & P_{12} & \dots \\ \vdots & \vdots & \vdots & \dots \\ P_{i0} & P_{i1} & P_{i2} & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix}$$

Now we have seen a one-step transition probability, we will define the k-step transition probability as P_{ij}^k. This probability states that a process in state i will be in state j after k additional transitions. So P_{ij}^k is defined as

$$P_{ij}^k \coloneqq \mathbb{P}(X_{n+k} = j \mid X_n = i)$$

for $k \geq 0$ and $i, j \geq 0$. The transition matrix for a k-step transition probability is denoted as $\mathbf{P}^k$. To compute this k-step transition probability, one can use the **Chapman-Kolmogorov equations**.

Proposition 2.1.1. For all $k, l \geq 0$ and $i, j \geq 0$, we have

$$P_{ij}^{k+l} = \sum_{n=0}^{\infty} P_{in}^k P_{nj}^l.$$

These equations are called the Chapman-Kolmogorov equations.

The probability $P_{in}^k P_{nj}^l$ represents that starting in state i the process will be in state n after k steps and from state n the process will go to state j in an additional l steps. We will now prove proposition (2.1.1).

Proof:

$$P_{ij}^{k+l} = \mathbb{P}(X_{k+l} = j \mid X_0 = i)$$

Now we apply the law of total probability. So:

$$\mathbb{P}(X_{k+l} = j \mid X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_{k+l} = j \cap X_k = n \mid X_0 = i)$$

$$= \sum_{n=0}^{\infty} \mathbb{P}(X_{k+l} = j \mid X_k = n, X_0 = i) \mathbb{P}(X_k = n \mid X_0 = i)$$

Using the Markov property results in

$$\sum_{n=0}^{\infty} \mathbb{P}(X_{k+l} = j \mid X_k = n, X_0 = i) \mathbb{P}(X_k = n \mid X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_{k+l} = j \mid X_k = n) \mathbb{P}(X_k = n \mid X_0 = i)$$

Which is by definition equal to the following:

$$\sum_{n=0}^{\infty} \mathbb{P}(X_{k+l} = j \mid X_k = n)\mathbb{P}(X_k = n \mid X_0 = i) = \sum_{n=0}^{\infty} P_{nj}^l P_{in}^k = \sum_{n=0}^{\infty} P_{in}^k P_{nj}^l$$

■

With the use of the Chapman-Kolmogorov equations, we can rewrite the transition matrix of $\mathbf{P}^{k+l}$ as a matrix multiplication of $\mathbf{P}^k$ and $\mathbf{P}^l$, so $\mathbf{P}^{k+l} = \mathbf{P}^k \cdot \mathbf{P}^l$.

So far, we have seen some definitions that are relevant for the Markov chain. In the next section, we will introduce some more terminology for the classification of states.

2.2 Classification of States

In this section, we will discuss the classification of states, the associated terminology and we will discuss some examples.
We will start with some terminology. We call a state j **accessible** from state i if $P_{ij}^k > 0$ for a $k \geq 0$, we denote this with $i \to j$. Two states i and j are said to **communicate** if they are accessible to each other, this is denoted by $i \leftrightarrow j$. The relation of communication satisfies the following three properties:

1. State i communicates with state i for all $i \geq 0$.
2. If state i communicates with state j, then communicates state j with state i.
3. If state i communicates with state j and state j communicates with state k, then communicates state i with state k.

Proof of property 1:
Property 1 is satisfied per definition, since we have $P_{ii}^0 = \mathbb{P}(X_0 = i \mid X_0 = i) = 1$.

■

Proof of property 2:
Property 2 follows immediately from the definition of communication.

■

Proof of property 3:
Assume that state i communicates with state j and state j communicates with state k. Now, we want to show that state i communicates with state k. There exists an n and m, both integers, such that $P_{ij}^n > 0$ and $P_{jk}^m > 0$. Because state i and state j are both accessible to each other, we have by definition of accessibility that $P_{ij}^n > 0$. The same holds true for state j and k. Now, we will apply the Chapman-Kolmogorov equations. So $P_{ik}^{n+m} = \sum_{r=0}^{\infty} P_{ir}^n P_{rk}^m \geq P_{ij}^n P_{jk}^m > 0$. Therefore, $P_{ik}^{n+m} > 0$. Hence, state k is accessible from state i. Similarly, we can show that state i is accessible from state k. So state i and state k communicate.

■

Two states that belong to the same **class** communicate with each other. Two classes of states can only be identical or disjoint as a consequence of the communication properties 1, 2 and 3. Further, we call a Markov chain **irreducible** if there is only one class. So this means that every state communicates with all of the other states. A state i is a **absorbing** state if $P_{ii} = 1$. This means that no other state is accessible from it. Lastly, we denote the probability that the process will reenter state i by f_i for any i with the assumption that the process started in state i, therefore f_i is denoted by $f_i = \mathbb{P}(\exists n \geq 1 : X_n = i \mid X_0 = i)$. And we call state i **recurrent** if $f_i = 1$ and **transient** if $f_i < 1$.

Assume that the process starts in state i and state i is recurrent, so $f_i = 1$. Then, by definition of a Markov Chain, the process will be starting over and over again when it reenters state i, hence state i will be reentered again. This argument can be repeated over and over again, which leads us to the following conclusion. If we start in state i and state i is recurrent, then the process will reenter state i repeatedly, in fact it will reenter state i infinitely often. Hence, the expected value of the number of entrances to state i is infinite.
Now, we assume that the process starts in state i and state i is transient, so $f_i < 1$ or $1 - f_i > 0$. So each time the process enters state i, there is a probability of $1 - f_i$ that state i will not be entered again. The probability of entering state i exactly n times is equal to $f_i^{n-1}(1 - f_i)$ for $n \geq 1$, which is equal to the geometric distribution with finite mean $\frac{1}{1-f_i}$. So we can summarize these paragraphs in the following proposition (2.2.1).

Proposition 2.2.1. State i is recurrent if and only if $\sum_{n=1}^{\infty} P_{ii}^n = \infty$ and state i is transient if and only if $\sum_{n=1}^{\infty} P_{ii}^n < \infty$

Proof:

Let the indicator function be defined as $\mathbb{I}_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i \end{cases}$.

So $\mathbb{I}_n$ is equal to one if the process at time n is in state i. Therefore, the summation of the indicator function, $\sum_{n=1}^{\infty} \mathbb{I}_n$, represents the number of periods of time that the process spends in state i. Hence, the conditional expectation of the number of visits to state i given the process starts in state i, $X_0 = i$, is given by:

$$\mathbb{E}[\sum_{n=1}^{\infty} \mathbb{I}_n \mid X_0 = i] = \sum_{n=1}^{\infty} \mathbb{E}[\mathbb{I}_n \mid X_0 = i] =$$

$$\sum_{n=1}^{\infty} (0 \cdot \mathbb{P}(X_n \neq i \mid X_0 = i) + 1 \cdot \mathbb{P}(X_n = i \mid X_0 = i)) =$$

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = i \mid X_0 = i) = \sum_{n=1}^{\infty} P_{ii}^n$$

Since the expected value of a recurrent state is equal to infinity and the expected value of a transient state is finite. We have that $\sum_{n=1}^{\infty} P_{ii}^n = \infty$ for a recurrent state i and $\sum_{n=1}^{\infty} P_{ii}^n < \infty$ if state i is transient.

■

Proposition 2.2.2. In a finite-state Markov chain not all states can be transient. In other words, in a finite-state Markov chain at least one state is recurrent.

Proof:

Consider a Markov chain with $p+1$ states, so $S = \{0, 1, 2, ..., p\}$. Assume that all states are transient. We will show that this assumption leads to a contradiction. If the process starts in state 0, then it might revisit state 0 several times, but after a finite amount of time T_0, state 0 will not be visited again. Same holds true if we start in state 1. Then the chain might revisit state 1 multiple times, but after a finite amount of time T_1, state 1 will not be revisited. We can give this argument for all the states $S = \{0, 1, 2, ..., p\}$, hence after a finite time $T = \max\{T_0, T_1, ..., T_p\}$ no states will be visited again. But the chain must be in some state after a finite amount of time T, hence a contradiction. So in a finite-state Markov chain, at least one state must be recurrent.

■

Corollary 2.2.1. If state i is recurrent and state i communicates with state j, then state j is also recurrent. In other words, if state i is recurrent and state i and state j are in the same class, then state j is also recurrent.

Proof:
State i communicates with state j. Hence, state i is accessible from state j, $P_{ji}^l > 0$ for some $l \in \mathbb{N}$, and state j is accessible from state i, $P_{ij}^k > 0$ for some $k \in \mathbb{N}$. For any $m \in \mathbb{N}$, we have

$$P_{jj}^{k+l+m} \geq P_{ji}^l P_{ii}^m P_{ij}^k \tag{2.1}$$

This is an application of the Chapman-Kolmogorov equations, which said $P_{ab}^{y+z} = \sum_{n=0}^{\infty} P_{an}^y P_{nb}^z$. The right-hand side of (2.1) is the probability that we go from state j to state j in $k+l+m$ steps via a path that goes from state j to state i in l steps, then from state i to state i in m steps and lastly from state i to state j in k steps. And the left-hand side of (2.1) is the probability that we go from state j to state j in $k+l+m$ steps.
Now, we take the summation over m. So

$$\sum_{m=0}^{\infty} P_{jj}^{k+l+m} \geq \sum_{m=0}^{\infty} P_{ji}^l P_{ii}^m P_{ij}^k = P_{ji}^l P_{ij}^k \sum_{m=0}^{\infty} P_{ii}^m$$

We know that $P_{ji}^l > 0$ and $P_{ij}^k > 0$, so $P_{ji}^l P_{ij}^k > 0$. And $\sum_{m=0}^{\infty} P_{ii}^m = \infty$, because state i is recurrent. Hence,

$$\sum_{m=0}^{\infty} P_{jj}^{k+l+m} = \infty.$$

So state j is recurrent.

■

Corollary 2.2.2. If state i is transient and state i communicates with state j, then state j is also transient. In other words, if state i is transient and state i and state j are in the same class, then state j is also transient.

Proof:
State i communicates with state j. Hence, state i is accessible from state j, $P_{ji}^l > 0$ for some $l \in \mathbb{N}$, and state j is accessible from state i, $P_{ij}^k > 0$ for some $k \in \mathbb{N}$. For any $m \in \mathbb{N}$, we have

$$P_{ii}^{k+l+m} \geq P_{ij}^k P_{jj}^m P_{ji}^l \tag{2.2}$$

This is an application of the Chapman-Kolmogorov equations, which said $P_{ab}^{y+z} = \sum_{n=0}^{\infty} P_{an}^y P_{nb}^z$. The right-hand side of (2.2) is the probability that we go from state i to state i in $k+l+m$ steps via a path that goes from state i to state j in k steps, then from state j to state j in m steps and lastly from state j to state i in l steps. And the left-hand side of (2.2) is the probability that we go from state i to state i in $k+l+m$ steps.
Now, we take the summation over m. So

$$\sum_{m=0}^{\infty} P_{ii}^{k+l+m} \geq \sum_{m=0}^{\infty} P_{ij}^k P_{jj}^m P_{ji}^l = P_{ij}^k P_{ji}^l \sum_{m=0}^{\infty} P_{jj}^m.$$

We know that $P_{ij}^k > 0$ and $P_{ji}^l > 0$, so $P_{ij}^k P_{ji}^l > 0$. Now, we divide both sides by $\frac{1}{P_{ij}^k P_{ji}^l}$. Thus,

$$\frac{1}{P_{ij}^k P_{ji}^l} \sum_{m=0}^{\infty} P_{ii}^{k+l+m} \geq \sum_{m=0}^{\infty} P_{jj}^m.$$

Hence,

$$\sum_{m=0}^{\infty} P_{jj}^{m} \leq \frac{1}{P_{ij}^{k} P_{ji}^{l}} \sum_{m=0}^{\infty} P_{ii}^{k+l+m}.$$

State i is transient, so $\sum_{m=0}^{\infty} P_{ii}^{k+l+m} < \infty$. Therefore,

$$\sum_{m=0}^{\infty} P_{jj}^{m} \leq \frac{1}{P_{ij}^{k} P_{ji}^{l}} \sum_{m=0}^{\infty} P_{ii}^{k+l+m} < \infty.$$

So state j is transient.

■

Hence, if states belong to the same class, then they are all recurrent or all transient.

The terminology, which is just discuss, will be applied in the following simple examples.

Example 2.2.1. Consider the Markov chain consisting of three states 0, 1, 2 and having transition matrix $\mathbf{P}$

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

This Markov chain is irreducible, which means that the Markov chain consist of only one class. And two states belong to the same class if they communicate with each other. So we will show that this holds true.
For example, it is possible to go from state 0 to state 2, since $0 \to 1 \to 2$.
State 1 is accessible from state 0 with $P_{01} = \frac{2}{3}$.
State 2 is accessible from state 1 with $P_{12} = \frac{1}{3}$.
It is also possible to go the other way around, since $2 \to 1 \to 0$.
State 1 is accessible from state 2 with $P_{21} = \frac{1}{2}$.
State 0 is accessible from state 1 with $P_{10} = \frac{1}{3}$.
So state 1 communicates with state 0 and state 1 communicates with state 2, so by property 3 of the communication relations we have that state 2 communicates with state 0. So they belong all to the same class. Hence, there is only one class. So the Markov chain is irreducible.

♦

Example 2.2.2. Consider the Markov chain consisting of four states 0, 1, 2, 3 and having transition matrix $\mathbf{P}$

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & P_{02} & P_{03} \\ P_{10} & P_{11} & P_{12} & P_{13} \\ P_{20} & P_{21} & P_{22} & P_{23} \\ P_{30} & P_{31} & P_{32} & P_{33} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$$

This Markov chain consists of the following three classes $\{0,1\}$, $\{2\}$ and $\{3\}$. We will explain how we concluded this.
$0 \to 0$ with probability $P_{00} = \frac{1}{3}$, $0 \to 1$ with probability $P_{01} = \frac{2}{3}$, $0 \to 2$ with probability $P_{02} = 0$ and $0 \to 3$ with probability $P_{03} = 0$.
$1 \to 0$ with probability $P_{10} = \frac{2}{3}$, $1 \to 1$ with probability $P_{11} = \frac{1}{3}$, $1 \to 2$ with probability $P_{12} = 0$ and $1 \to 3$ with probability $P_{13} = 0$.
So state 1 is accessible from state 0 and state 0 is accessible from state 1, hence state 0 and state 1 communicate. So state 0 and state 1 are in the same class, $\{0,1\}$.
$2 \to 0$ with probability $P_{20} = 0$, $2 \to 1$ with probability $P_{21} = 0$, $2 \to 2$ with probability $P_{22} = 1$ and $2 \to 3$ with probability $P_{23} = 0$.
So state 2 is an absorbing state, no other state is accessible from it.

Hence, state 2 has its own class, $\{2\}$.
$3 \to 0$ with probability $P_{30} = \frac{1}{2}$, $3 \to 1$ with probability $P_{31} = \frac{1}{4}$, $3 \to 2$ with probability $P_{32} = \frac{1}{8}$ and $3 \to 3$ with probability $P_{33} = \frac{1}{8}$.
So every state is accessible from state 3, but the reverse is not true, hence $\{3\}$.
Therefore, the classes of the Markov chain are $\{0,1\}$, $\{2\}$ and $\{3\}$.

♦

Example 2.2.3. This example is a sequel of example 2.2.1. So we have a Markov chain consisting of three states 0, 1, 2, with transition matrix $\mathbf{P}$

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

We have seen that this Markov chain is irreducible, hence it consist of just one class. So all the states communicate with each other. And since it is a finite chain, all states must be recurrent.

♦

2.3 Limiting Probabilities

In this section, we will consider additional properties of the states of a Markov chain.
Again, we will start with some terminology. A state i is said to have **period** d if $P_{ii}^k = 0$, whenever k can not be divided by d. Where d is the largest integer with this property. This means, for example, that if the process can enter state i only at the times 3, 6, 9, ... then state i has period 3. We call a state i **aperiodic** if state i has period 1.

Proposition 2.3.1. If state i has period d and state i communicates with state j, then state j has also period d. In other words, if state i has period d and state i and state j are in the same class, then state j has also period d.

Proof:
Define d_i as the period of state i and d_j as the period of state j. State i communicates with state j. Hence, state i is accessible from state j, $P_{ji}^l > 0$ for some $l \in \mathbb{N}$, and state j is accessible from state i, $P_{ij}^k > 0$ for some $k \in \mathbb{N}$. Then, we have

$$P_{ii}^{k+l} = \sum_{n=0}^{\infty} P_{in}^k P_{ni}^l$$

This is an application of the Chapman-Kolmogorov equations, which said $P_{ab}^{y+z} = \sum_{n=0}^{\infty} P_{an}^y P_{nb}^z$.
So,

$$P_{ii}^{k+l} = \sum_{n=0}^{\infty} P_{in}^k P_{ni}^l \geq P_{ij}^k P_{ji}^l > 0.$$

Therefore, d_i, the period of state i, divides $k+l$. For some $m \in \mathbb{N}$, we have

$$P_{ii}^{k+l+m} \geq P_{ij}^k P_{jj}^m P_{ji}^l$$

Which is again an application of the Chapman-Kolmogorov equations.
Now, we have that d_i, the period of state i, divides $k+l+m$.
So, d_i divides $k+l$ and d_i divides $k+l+m$.
Hence, there exist a x and a y both an integer, such that $xd_i = k+l$ and $yd_i = k+l+m$.
Therefore, $m = (y-x)d_i$. Which implies that d_i divides m, with m such that $P_{jj}^m > 0$.

So d_j is the largest divisor of m, by definition of period.
Therefore, $d_i \leq d_j$.
Hence, by symmetry we have $d_j \leq d_i$.
Thus $d_i = d_j$. And therefore state j has the same period as state i.

■

State i is said to be **positive recurrent** if the process starts in state i, state i is recurrent and the expected time until the process returns to state i is finite. This is denoted by $\mathbb{E}[\tau_{ii} \mid X_0 = i] < \infty$, with $\tau_{ii} \coloneqq \min\{\exists n \geq 1 : X_n = i \mid X_0 = i\}$. It can be shown that in a finite-state Markov chain all recurrent states are positive recurrent. States that are positive recurrent and aperiodic are called **ergodic**. We will state the next theorem without a proof.

Theorem 2.3.1. If a state is ergodic, then

$$\lim_{r\to\infty} \frac{\sum_{m=0}^{r} P_{ii}^{m}}{r+1} \coloneqq \frac{1}{\mathbb{E}[\tau_{ii} \mid X_0 = i]} \quad \text{with} \quad \mathbb{E}[\tau_{ii} \mid X_0 = i] > 0.$$

In other words, state i is said to be positive recurrent if

$$\lim_{r\to\infty} \frac{\sum_{m=0}^{r} P_{ii}^{m}}{r+1} > 0 \quad \text{and} \quad \lim_{r\to\infty} \frac{\sum_{m=0}^{r} P_{ii}^{m}}{r+1} < \infty$$

in case both limits exists.

Proposition 2.3.2. If state i is positive recurrent and state i communicates with state j, then state j is also positive recurrent. In other words, if state i is positive recurrent and state i and state j are in the same class, then state j is also positive recurrent.

Proof:
State i communicates with state j. Hence, state i is accessible from state j, $P_{ji}^{l} > 0$ for some $l \in \mathbb{N}$, and state j is accessible from state i, $P_{ij}^{k} > 0$ for some $k \in \mathbb{N}$. For any $m \in \mathbb{N}$, we have

$$P_{jj}^{k+l+m} \geq P_{ji}^{l} P_{ii}^{m} P_{ij}^{k}$$

This is an application of the Chapman-Kolmogorov equations, which said $P_{ab}^{y+z} = \sum_{n=0}^{\infty} P_{an}^{y} P_{nb}^{z}$. Now, we take the summation over m. So

$$\sum_{m=0}^{\infty} P_{jj}^{k+l+m} \geq \sum_{m=0}^{\infty} P_{ji}^{l} P_{ii}^{m} P_{ij}^{k} = P_{ji}^{l} P_{ij}^{k} \sum_{m=0}^{\infty} P_{ii}^{m}.$$

Further,

$$\lim_{t\to\infty} \frac{\sum_{m=0}^{t} P_{jj}^{k+l+m}}{t+1} \geq P_{ji}^{l} P_{ij}^{k} \lim_{r\to\infty} \frac{\sum_{m=0}^{r} P_{ii}^{m}}{r+1}$$

We know that $P_{ji}^{l} > 0$ and $P_{ij}^{k} > 0$, so $P_{ji}^{l} P_{ij}^{k} > 0$. And $\lim_{r\to\infty} \frac{\sum_{m=0}^{r} P_{ii}^{m}}{r+1} > 0$, because state i is positive recurrent. Hence,

$$\lim_{t\to\infty} \frac{\sum_{m=0}^{t} P_{jj}^{k+l+m}}{t+1} > 0.$$

So state j is positive recurrent.

■

The probability distribution $\pi = (\pi_0, \pi_1, \pi_2, ...)$ is called a stationary or limiting distribution if we have

$$\pi \mathbf{P} = \pi. \tag{2.3}$$

By multiplying both sides of equation (2.3) by $\mathbf{P}$, we obtain the following:

$$\pi \mathbf{P}^2 = \pi \mathbf{P}.$$

Hence,

$$\pi \mathbf{P}^2 = \pi \mathbf{P} = \pi,$$

by equation (2.3). Therefore, by induction we have

$$\pi \mathbf{P}^n = \pi,$$

with π the initial distribution.

We will state the next theorem without a proof. The proof of this theorem can be found on page 111-113 of ([5]).

Theorem 2.3.2. For an irreducible ergodic Markov chain $\lim_{k\to\infty} P_{ij}^k$ exists and is independent of i. Furthermore, letting $\pi_j = \lim_{k\to\infty} P_{ij}^k$ with $j \geq 0$ be the limiting probabilities, then π_j is the unique nonnegative solution of $\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}$ with $j \geq 0$ and $\sum_{j=0}^{\infty} \pi_j = 1$.

Now, we will discuss the following example.

Example 2.3.1. Assume we have the following transition matrix $\mathbf{P}$

$$\mathbf{P} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$$

Then the limiting probabilities π_i satisfy

$$\pi_0 = 0.3\pi_0 + 0.6\pi_1$$

$$\pi_1 = 0.7\pi_0 + 0.4\pi_1$$

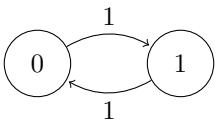
$$\pi_0 + \pi_1 = 1$$

Hence, $\pi_0 = \frac{6}{13}$ and $\pi_1 = \frac{7}{13}$.

♦

We have seen many different terminology for the states of a Markov chain and we will discuss these different types of Markov chains in the following examples.

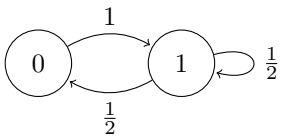
Example 2.3.2. Periodic: Consider the Markov chain consisting of two states 0 and 1.



This Markov chain has period $d = 2$. If we start in state 0, then we are back at state 0 in two time steps. The same holds true for state 1. Thus, starting from state 0, we only return to state 0 at times $n = 2, 4, 6, ...$

♦

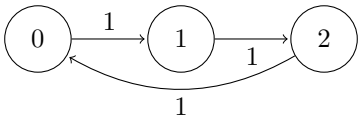
Example 2.3.3. Aperiodic: Consider the Markov chain consisting of two states 0 and 1.



This Markov chain is aperiodic, since $P_{11} = \frac{1}{2}$. Hence, we can go to state 0 in two steps or in more steps, because state 1 has the opportunity to revisit state 1 immediately.

♦

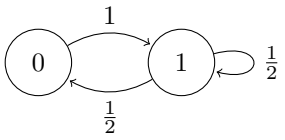
Example 2.3.4. Recurrent and Positive recurrent: Consider the Markov chain consisting of three states 0, 1 and 2.



All states in this Markov chain are recurrent and it is a finite-state Markov chain, hence all states are positive recurrent.

♦

Example 2.3.5. Ergodic: Consider the Markov chain consisting of two states 0 and 1.



We have seen that this Markov chain is aperiodic in an example above. Further, the states in this chain are recurrent and it is a finite-state Markov chain. Therefore, the states are positive recurrent. Hence, this chain is ergodic.

♦

Since we have discussed Markov chains thoroughly, we can move on to a more advanced topic, namely Markov decision processes. This will be discussed it the next chapter.

Chapter 3

Markov Decision Processes

In this chapter, we will discuss the theory behind discrete-time Markov decision processes. This will be done in three sections, namely the model formulation, finite-horizon and infinite-horizon.

3.1 Model Formulation

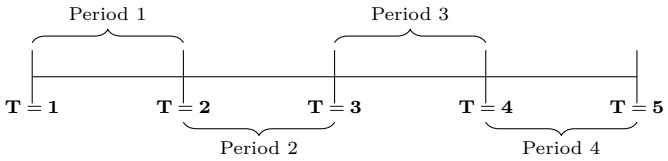
In this section, we will introduce the basic components of a discrete-time Markov decision process. We will discuss the formulation of the Markov decision process model in detail. As we have seen in chapter one, a Markov decision process consists of five elements: decision epochs or times, states, actions, rewards, and transition probabilities. Besides the decision epochs, states, actions, rewards and transition probabilities, we have a decision maker that observes the process and may select actions at each decision epoch to influence the system and gain rewards. Mathematically, we can formulate a Markov decision process by the collection of the five elements

$$\{T, S, A_s, p_t(\cdot \mid s,a), r_t(s,a) \mid t \in T, s \in S, a \in A_s\}.$$

The five elements and its notation will be explained below. After that, we will discuss the decision rules and policies, which we already saw in chapter one as well. We will end this section with an example of a Markov decision process. We will now start with the explanation of the components of a discrete-time Markov decision process.

Decision epochs or **decision times** are given points in time, where decisions are made by the decision maker. We denote T as the set of decision times and we assume that the set T is discrete. We make this assumption, because we are only interested in discrete-time Markov decision processes. The set of decision times T can either be finite or infinite. When the set of decision times is finite, we denote $T = \{1, 2, ..., N\}$ with $N < \infty$. In the case that T is infinite, we have $T = \{1, 2, ...\}$. Further, we denote the elements of the set T as t, which we refer to as time t. Also, we call the decision problem a **finite-horizon** problem, if N is finite. And a decision problem is an **infinite-horizon** problem, if N is infinite. These problems will be discussed in section two and section three of this chapter. We assumed that our set of decision times T is discrete, and therefore the decisions will be made at all decision times. Further, time will be divided into **periods** or **stages**. Our model will be formulated such that each decision epoch corresponds to the beginning of a period. So the last decision will be made at period $N-1$ if our set of decision times $T = \{1, 2, ..., N\}$. Therefore, we we will call the problem an $N-1$ period problem. In the following example, we will explain the $N-1$ period problem graphically.

Example 3.1.1. In this example, we have the set of decision times T defined as $T = \{1,2,3,4,5\}$. So at each time t with t an element of T, the decision maker makes a decision. Further, we defined that a period or stage will start at the beginning of each decision time T, hence we have 4 stages. Thus, this is a 4-period problem.



♦

Secondly, we will explain the state and actions set. The process occupies a **state** at each decision time, we denote the set of all possible states by S. The elements of S will be denoted by s. When the decision maker observes the system in state $s \in S$ at a given decision time t, then he may choose an **action** $a \in A_s$, where A_s is the set of possible actions in that specific state s. Further, we denote the set of all possible actions by $A = \bigcup_{s \in S} A_s$. Also, actions may be chosen either deterministically or randomly. If actions are chosen at random, then the collection of probability distributions is denoted by $\mathcal{P}(A_s)$. This means that the decision maker may select an action $a \in A_s$ with probability $q(a)$, where $q(\cdot) \in \mathcal{P}(A_s)$. This will be further explained when we will discuss decision rules.

Lastly, we will discuss the rewards and the transition probabilities. Choosing an action $a \in A_s$ in state $s \in S$ at decision time t leads to the following two results, namely:

1. The decision maker receives a **reward** $r_t(s,a)$
2. The system evolves to the next state according to the **probability distribution** $p_t(s,a)$

Therefore, the reward function $r_t(s,a)$ denotes the value of the reward received at time $t \in T$ for $s \in S$ and $a \in A$. This value can be positive as well as negative. If the reward depends on the state j, with state j the state of the next decision time, then we denote the value of the reward received at time $t \in T$ for $s \in S$ and $a \in A$ by $r_t(s,a,j)$. So our reward function $r_t(s,a)$ may be computed in the following way, where we assume that $\sum_{j \in S} p_t(j \mid s,a) = 1$:

$$r_t(s,a) = \sum_{j \in S} r_t(s,a,j) p_t(j \mid s,a) \tag{3.1}$$

The **transition probability function** $p_t(j \mid s,a)$ denotes the probability that the system will be in state j at the next decision epoch $t+1$, when the decision maker chooses an action $a \in A_s$ in state $s \in S$ at decision time t.

We have discussed the components of a discrete-time Markov decision process model. So we can conclude that the Markov decision process can be formulated by the collection of the five elements, hence

$$\{T, S, A_s, p_t(\cdot \mid s,a), r_t(s,a) \mid t \in T, s \in S, a \in A_s\}.$$

Now, we will explain decision rules and policies. A **decision rule** prescribes a procedure for choosing an action in each state at a given decision epoch. We can identify four classes of decision rules, namely

1. Markovian and deterministic, denoted by MD
2. History dependent and deterministic, denoted by HD
3. Markovian and randomized, denoted by MR
4. History dependent and randomized, denoted by HR

These distinctions are made depending on how the decision rules incorporated past information and how the actions are selected. We will explain each class individually, started with the first class. Markovian and deterministic decision rules are functions $d_t : S \to A_s$, which specifies the choice of action when the process occupies state $s \in S$ at decision time t, hence for each $s \in S$, $d_t(s) \in A_s$. Decision rules, that are specified by this function, are said to be **Markovian**, because it depends only on the current state and not on the past states, so it is memoryless. And these decision rules are called **deterministic**, because it chooses an action with certainty. So the states are not chosen at random, but the course of the states are predetermined. A deterministic decision rule is said to be **history dependent** if it depends on the past states and actions, so it is not memoryless. Therefore, d_t is a function of the history $h_t = (s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t)$, with s_i and a_i the state and action at decision time $t = i$. Further, the history h_t follows a recursion $h_t = (h_{t-1}, a_{t-1}, s_t)$. In general, if we use a decision rule that is history dependent the the decision maker observes h_t and chooses actions from the set A_{s_t}. The set of all histories h_t is denoted by H_t. Because h_t follows a recursion $h_t = (h_{t-1}, a_{t-1}, s_t)$, we have that

$$H_1 = S,$$

$$H_2 = H_1 \times A \times S = S \times A \times S,$$

$$H_3 = H_2 \times A \times S = S \times A \times S \times A \times S,$$

and so on. Hence, $H_t = H_{t-1} \times A \times S$. Note that this is a product of sets. Thus, history dependent and deterministic decision rules are functions $d_t : H_t \to A$, with the restriction that $d_t(h_t) \in A_{s_t}$. Markovian and randomized decision rules are functions $d_t : S \to \mathcal{P}(A)$, where $\mathcal{P}(A)$ denotes the collection of probability distributions as we have seen before. A decision rule d_t that is **randomized** specifies a probability distribution $q_{d_t}(\cdot)$ on the set of actions. In the case of a Markovian and randomized decision rule, we have $q_{d_t(s_t)}(\cdot) \in \mathcal{P}(A_{s_t})$. A history dependent and randomized decision rule is a function $d_t : H_t \to \mathcal{P}(A)$, with $q_{d_t(h_t)}(\cdot) \in \mathcal{P}(A_{s_t})$ for all $h_t \in H_t$. Note that a deterministic decision rule is a specific case of a randomized decision rule, because we can take $q_{d_t(s_t)}(a) = 1$ or $q_{d_t(h_t)}(a) = 1$ for some $a \in A_s$. This means that with absolute certainty the next states are determined. Because the actions are chosen non-randomly. We have seen the different classes of decision rules. The **set of decision rules** at time t is denoted by D_t^K, with $K \in \{MD, HD, MR, HR\}$. All this information is summarized at table (3.1).

Table 3.1: Classes of Decision Rules

Action Choice		
History Dependence	Deterministic	Randomized
Markovian	$d_t(s_t) \in A_{s_t}$, D_t^{MD}	$q_{d_t(s_t)}(\cdot) \in \mathcal{P}(A_{s_t})$, D_t^{MR}
History Dependent	$d_t(h_t) \in A_{s_t}$, D_t^{HD}	$q_{d_t(h_t)}(\cdot) \in \mathcal{P}(A_{s_t})$, D_t^{HR}

Under the four classes of decision rules, the rewards and transition probability become functions on S or H_t depending on the class. For a Markovian and deterministic decision rule $d_t \in D_t^{MD}$, the reward equals $r_t(s, d_t(s))$ and the transition probability equals $p_t(j \mid s, d_t(s))$. For a history dependent and deterministic decision rule $d_t \in D_t^{HD}$, we have that the reward equals $r_t(s, d_t(h_t))$

and the transition probability $p_t(j \mid s, d_t(h_t))$, whenever $h_t = (h_{t-1}, a_{t-1}, s_t)$.
For the randomized decision rules it becomes a bit different. If the decision rule is Markovian and randomized $d_t \in D_t^{MR}$, then the expected reward satisfies

$$r_t(s, d_t(s)) = \sum_{a \in A_s} r_t(s, a) q_{d_t(s_t)}(a) \tag{3.2}$$

and the transition probability satisfies

$$p_t(j \mid s, d_t(s)) = \sum_{a \in A_s} p_t(j \mid s, a) q_{d_t(s_t)}(a). \tag{3.3}$$

For a decision rule that is history dependent and randomized $d_t \in D_t^{HR}$, we have that the expected reward satisfies

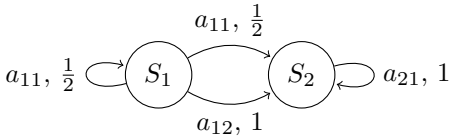
$$r_t(s, d_t(h_t)) = \sum_{a \in A_s} r_t(s, a) q_{d_t(h_t)}(a) \tag{3.4}$$

and the transition probability satisfies

$$p_t(j \mid s, d_t(h_t)) = \sum_{a \in A_s} p_t(j \mid s, a) q_{d_t(h_t)}(a). \tag{3.5}$$

Lastly, we will discuss policies. A **policy** π provides the decision maker with a prescription of how to choose actions under any possible future state or history. So a policy π is a sequence of decision rules, $\pi = (d_1, d_2, ..., d_{N-1})$ where $d_t \in D_t^K$ for $t = 1, 2, ..., N-1$. The set of all policies of a class $K \in \{MD, HD, MR, HR\}$ is denoted by $\prod^K$, with $\prod^K = D_1^K \times D_2^K \times ... \times D_{N-1}^K$. Further, a policy is said to be **stationary** if $d_t = d$ for all $t \in T$. Hence, a stationary policy has the following form $\pi = (d, d, ...)$, which we will denote by d^∞. We will come back to the stationary policies, when we discuss the infinite-horizon Markov decision processes. In the following example, we will discuss a two state Markov decision process.

Example 3.1.2. A Two State Markov Decision Process
Consider the following representation of the two state Markov decision process.



In this example, we assume that the rewards and transition probabilities are the same at each epoch. There are two states $S = \{S_1, S_2\}$. In state S_1, the decision maker chooses either action a_{11} or action a_{12}. In state S_2 the only choice the decision maker has is action a_{21}. Choosing action a_{11} in state S_1 leads to an immediate reward of five units, and the system will evolve to state S_1 with a probability of $\frac{1}{2}$ and to state S_2 with a probability of $\frac{1}{2}$ as well. If the decision maker chooses action a_{12} in state S_1, then he will receive an immediate reward of ten units and the system evolves to state S_2 with a probability of 1. In state S_2, the decision maker has no other choice than to choose action a_{21}, by doing so he will receive an immediate reward of minus one unit and the system stays in state S_2 with probability 1.

So formally, our model is formulated in the following way:

- Decision epochs: $T = \{1, 2, ..., N\}$, $N < \infty$
- States: $S = \{S_1, S_2\}$
- Actions: $A_{s_1} = \{a_{11}, a_{12}\}$ and $A_{s_2} = \{a_{21}\}$
- Rewards:
  $r_t(s_1, a_{11}) = 5$, $r_t(s_1, a_{12}) = 10$, $r_t(s_2, a_{21}) = -1$ for $t \in \{1, 2, ...N-1\}$
  $r_N(s_1) = 0$ and $r_N(s_2) = 0$ for $t = N$
- Transition probabilities:
  $p_t(s_1 \mid s_1, a_{11}) = \frac{1}{2}$, $p_t(s_1 \mid s_1, a_{12}) = 0$, $p_t(s_1 \mid s_2, a_{21}) = 0$,
  $p_t(s_2 \mid s_1, a_{11}) = \frac{1}{2}$, $p_t(s_2 \mid s_1, a_{12}) = 1$, $p_t(s_2 \mid s_2, a_{21}) = 1$.

Suppose for a moment that the rewards corresponding to action a_{11} depend upon the state at the next epoch. Let $r_t(s_1, a_{11}, s_1) = 2$ and $r_t(s_1, a_{11}, s_2) = 8$. Then the expected reward $r_t(s, a)$ will be equal to 5, because of equation (3.1). Since

$$r_t(s_1, a_{11}) = \sum_{j \in S} r_t(s, a, j) p_t(j \mid s, a) =$$

$$r_t(s_1, a_{11}, s_1) p_t(s_1 \mid s_1, a_{11}) + r_t(s_1, a_{11}, s_2) p_t(s_2 \mid s_1, a_{11}) =$$

$$2 \times \frac{1}{2} + 8 \times \frac{1}{2} = 5.$$

Now, we will continue our example by providing some of the policies which were discussed in the section above. We will only elaborate on the Markovian policies and we assume that $N = 3$, which implies that the decisions are only made at decision times 1 and 2. Therefore, our policies are represented as $\pi^K = (d_1^K, d_2^K)$ with $K = \{MD, MR\}$. The first policy we will discuss is the Markovian and deterministic policy π^{MD}.

- Decision epoch 1: $d_1^{MD}(s_1) = a_{11}$ and $d_1^{MD}(s_2) = a_{21}$
- Decision epoch 2: $d_2^{MD}(s_1) = a_{12}$ and $d_2^{MD}(s_2) = a_{21}$

This means that the decision maker chooses in the first decision epoch action a_{11} if he is in state S_1 and action a_{21} if he is in state S_2. In the second decision epoch he chooses action a_{12} in state S_1 and again action a_{21} if he is in state S_2.

The other policy we will discuss is the Markovian and randomized policy π^{MR}. We assume that the probability of choosing action a_{11} is equal to $\frac{4}{5}$ in the first decision epoch and equal to $\frac{2}{5}$ in the second decision epoch. Further, we assume that the probability of choosing action that the decision maker chooses a_{12} is equal to $\frac{1}{5}$ in the first decision epoch and equal to $\frac{3}{5}$ in the second one. And lastly, the decision maker chooses with probability of 1 action a_{21} in both decision times. So our policy will look like

- Decision epoch 1: $q_{d_1^{MR}(s_1)}(a_{11}) = \frac{4}{5}$, $q_{d_1^{MR}(s_1)}(a_{12}) = \frac{1}{5}$ and $q_{d_1^{MR}(s_2)}(a_{21}) = 1$
- Decision epoch 2: $q_{d_2^{MR}(s_1)}(a_{11}) = \frac{2}{5}$, $q_{d_2^{MR}(s_1)}(a_{12}) = \frac{3}{5}$ and $q_{d_2^{MR}(s_2)}(a_{21}) = 1$

♦

Before starting the next section, we will provide some additional notation for a Markov decision process, which is needed for the understanding of finite- and infinite-horizon Markov decision processes. Throughout this discussion, we will assume that the set of states S and the set of actions A are both discrete and the set of decision times is given by $T = \{1, 2, ..., N\}$ with $N \leq \infty$. In general, a probability space consists of three components, namely a sample space Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω and a probability measure $\mathbb{P}$. Thus the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability

space. We assume in this model that the probability space consists of a sample space Ω, a σ-algebra of Borel measurable subsets of Ω, $\mathcal{B}(\Omega)$ and a probability measure $\mathbb{P}$ on $\mathcal{B}(\Omega)$. We will not further elaborate on these concepts, we will just use the notation.
In a finite-horizon Markov decision process, which will be explained in dept in section two of this chapter, we choose the sample space Ω as

$$\Omega = S \times A \times S \times ... \times A \times S = (S \times A)^{N-1} \times S$$

and an element of Ω is denoted by ω. So for $\omega \in \Omega$, we have $\omega = (s_1, a_1, s_2, a_2, ...a_{N-1}, s_N)$, which we refer to as a **sample path**. In an infinite-horizon Markov decision process, which we will elaborate on in section three of this chapter, we choose the sample space Ω as $\Omega = (S \times A)^{\infty}$ and for $\omega \in \Omega$, we have $\omega = (s_1, a_1, s_2, a_2, ...)$.
Further, we define the random variable X_t as the state at time $t \in T$, the random variable Y_t is defined as the action at time t for $t \in T$ and the random variable Z_t is the history process at time $t \in T$, when the observed sequence of states and actions is ω. Therefore, we have $X_t(\omega) = s_t$, $Y_t(\omega) = a_t$, $Z_1(\omega) = s_1$ and $Z_t(\omega) = (s_1, a_1, ..., s_t)$.
We denote the **initial distribution** of the state by the probability distribution $P_1(\cdot)$, most times we will assume that $P_1(s_1) = 1$ for some $s_1 \in S$. If we have a history dependent and randomized policy $\pi = (d_1, d_2, ..., d_{N-1})$, then the probability P^{π} on our measurable space $(\Omega, \mathcal{B}(\Omega))$ is induced through the following probabilities for $t \in T$

$$P^{\pi}(\{X_1 = s\}) = P_1(s), \tag{3.6}$$

$$P^{\pi}(\{Y_t = a \mid Z_t = h_t\}) = q_{d_t(h_t)}(a), \tag{3.7}$$

$$P^{\pi}(\{X_t = s \mid Z_t = (h_{t-1}, a_{t-1}, s_t), Y_t = a_t\}) = p_t(s \mid s_t, a_t), \tag{3.8}$$

such that the probability of a sample path $\omega = (s_1, a_1, s_2, a_2, ...a_{N-1}, s_N)$ is given by

$$P^{\pi}(\omega) = P^{\pi}(s_1, a_1, s_2, a_2, ...a_{N-1}, s_N) =$$

$$P_1(s) q_{d_1(s_1)}(a_1) p_1(s_2 \mid s_1, a_1) q_{d_2(h_2)}(a_2) \dots q_{d_{N-1}(h_{N-1})}(a_{N-1}) p_{N-1}(s_N \mid s_{N-1}, a_{N-1}). \tag{3.9}$$

The last equation (3.9) can be simplified if the policy π is deterministic, so it can either be history dependent and deterministic or Markovian and deterministic. In this case the expression will be

$$P^{\pi}(\omega) = P^{\pi}(s_1, a_1, s_2, a_2, ...a_{N-1}, s_N) = P_1(s) p_1(s_2 \mid s_1, a_1) \dots p_{N-1}(s_N \mid s_{N-1}, a_{N-1}), \tag{3.10}$$

because $q_{d_1(s_1)}(a_1) = 1$ and $q_{d_t(h_t)}(a) = 1$ for all $a \in A_s$ and $t \in T$. Further, for a non-deterministic policy, we have the following expression for $t \in T$

$$P^{\pi}(s_1, a_1, s_2, a_2, ...a_{t-1}, s_t) =$$

$$P_1(s) q_{d_1(s_1)}(a_1) p_1(s_2 \mid s_1, a_1) q_{d_2(h_2)}(a_2) \dots q_{d_{t-1}(h_{t-1})}(a_{t-1}) p_{t-1}(s_t \mid s_{t-1}, a_{t-1}) \tag{3.11}$$

Therefore, we can calculate for the non-deterministic policy the conditional probability

$$P^{\pi}(a_t, s_{t+1}, a_{t+1}, ...a_{N-1}, s_N \mid s_1, a_1, ...a_{t-1}, s_t)$$

by division of equation (3.9) and equation (3.11), hence

$$P^{\pi}(a_t, s_{t+1}, a_{t+1}, ...a_{N-1}, s_N \mid s_1, a_1, ...a_{t-1}, s_t) = \frac{P^{\pi}(s_1, a_1, s_2, a_2, ...a_{N-1}, s_N)}{P^{\pi}(s_1, a_1, s_2, a_2, ...a_{t-1}, s_t)} \tag{3.12}$$

The expression (3.12) simplifies to

$$P^{\pi}(a_t, s_{t+1}, a_{t+1}, ...a_{N-1}, s_N \mid s_1, a_1, ...a_{t-1}, s_t) =$$

$$q_{d_t(h_t)}(a_t)p_t(s_{t+1} \mid s_t, a_t)q_{d_{t+1}(h_{t+1})}(a_{t+1}) \dots q_{d_{N-1}(h_{N-1})}(a_{N-1})p_{N-1}(s_N \mid s_{N-1}, a_{N-1}) \quad (3.13)$$

In a similar way, we have that the conditional probability

$$P^{\pi}(a_t, s_{t+1}, a_{t+1}, ...a_{N-1}, s_N \mid s_1, a_1, ...a_{t-1}, s_t)$$

for a Markovian policy is given by

$$P^{\pi}(a_t, s_{t+1}, a_{t+1}, ...a_{N-1}, s_N \mid s_1, a_1, ...a_{t-1}, s_t) = P^{\pi}(a_t, s_{t+1}, a_{t+1}, ...a_{N-1}, s_N \mid s_t), \quad (3.14)$$

because d_t depends only on the current state of the process and not on the past states since it satisfies the Markov property. Lastly, the **Markov reward process** is defined as a stochastic process $\{(X_t, r_t(X_t, Y_t)) \mid t \in T\}$, when the policy π is Markovian. The first component of the reward process, X_t, represents the state of the process at time t. The second component, $r_t(X_t, Y_t)$, represents the reward received in state X_t at time t, when action Y_t is used.

In the next section, we will discuss finite-horizon Markov decision processes.

3.2 Finite-Horizon Markov Decision Processes

In this section, we will explain discrete-time finite-horizon Markov decision processes. We will introduce the following concepts; optimality criteria, optimal policies and optimality equations. We will start with the optimality criteria. As we have seen before, the decision maker receives rewards in each decision epoch $T = \{1, 2, ..., N\}$, $N < \infty$. Since the rewards are unknown prior to choosing a policy, the decision maker must observe the sequence of rewards as random, which we denote as $R = (R_1, R_2, ..., R_N)$. The set of all possible reward sequences is denoted by $\mathcal{R}$. Further, once a policy is selected it induces a probability distribution $P^{\pi}_{\mathcal{R}}(\cdot)$ on $\mathcal{R}$, which is defined as

$$P^{\pi}_{\mathcal{R}}(\rho_1, \rho_2, ..., \rho_N) \coloneqq P^{\pi}(\{(s_1, a_1, ...a_{N-1}, s_N) \mid (r_1(s_1, a_1), ..., r_N(s_N)) = (\rho_1, \rho_2, ..., \rho_N)\})$$

So the goal is to choose a policy that corresponds to a sequence of rewards that is most rewarding for the decision maker. Thus we need to compare the different policies based on the decision makers preferences for the different sequences of rewards and the probability in which these rewards sequences occur. Hence, we need to develop methods to compare sequences of rewards with each other. The first way to compare reward sequences is using a **utility function** $\Psi : \mathbb{R}^N \to \mathbb{R}$. The utility function has the property that $\Psi(u) \geq \Psi(v)$, whenever the decision maker prefers the sequence of rewards $u = (u_1, u_2, ..., u_N)$ over the sequence of rewards $v = (v_1, v_2, ..., v_N)$. If the decision maker does not prefer the sequence of rewards $u = (u_1, u_2, ..., u_N)$ over the sequence of rewards $v = (v_1, v_2, ..., v_N)$, then $\Psi(u) \leq \Psi(v)$. If he does not favour one over the other, then we have that $\Psi(u) = \Psi(v)$. Another way, is to use the **expected utility** of policy π, which is defined as

$$\mathbb{E}^{\pi}[\Psi(R)] \coloneqq \sum_{(\rho_1, ..., \rho_N) \in \mathcal{R}} \Psi(\rho_1, ..., \rho_N) P^{\pi}_{\mathcal{R}}(\rho_1, ..., \rho_N), \quad (3.15)$$

and we will assume that $\Psi(\rho_1, ..., \rho_N) = \sum_{i=1}^{N} \rho_i$. In that case, if the decision maker favours policy π over policy ν, then the **expected utility criterion** is given by $\mathbb{E}^{\pi}[\Psi(R)] \geq \mathbb{E}^{\nu}[\Psi(R)]$ for the reward sequence $R = (R_1, R_2, ..., R_N)$. If the policies are equivalent and the decision maker does not favour any policy over the other, then we have $\mathbb{E}^{\pi}[\Psi(R)] = \mathbb{E}^{\nu}[\Psi(R)]$. Another method of comparing is the **expected total reward criterion**. This criterion states that the decision maker should favour the policy π with the highest **expected total reward**, which is denoted by v^{π}_N. The expected total reward for a history dependent and randomized policy π with state s as

initial state is defined as

$$v_N^\pi(s) \coloneqq \mathbb{E}_s^\pi[\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N)]. \tag{3.16}$$

If we have a history dependent and deterministic policy π, then we can express the expected total reward as

$$v_N^\pi(s) \coloneqq \mathbb{E}_s^\pi[\sum_{t=1}^{N-1} r_t(X_t, d_t(h_t)) + r_N(X_N)], \tag{3.17}$$

since we have $Y_t = d_t(h_t)$ for each decision epoch. This criterion assumes that a unit reward is received in each of the N decision periods is equal or less valuable than a sequence of rewards in which all N units of rewards are received in the first or last decision period. In other words, the timing of rewards is insignificant for the decision maker. However, if the timing of receiving rewards becomes significant for the decision maker, because the value of reward depends on when it is received, then we can use a discount factor. The **discount factor** is a scalar λ, $0 \leq \lambda < 1$, which determines the value at time t of a one unit reward received at time $t+1$. For a history dependent and randomized policy π, the **expected total discounted reward** is defined by

$$v_{N,\lambda}^\pi(s) \coloneqq \mathbb{E}_s^\pi[\sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t, Y_t) + \lambda^{N-1} r_N(X_N)]. \tag{3.18}$$

When we start with the infinite-horizon Markov decision processes, then discounting will play a major part. This will be explained further in section three of this chapter. Now, we will discuss optimal history dependent and randomized policies more specifically. So in a Markov decision process, we are interested in finding a policy π^* which has the largest expected total reward and determining the value of that expected total reward. Hence, we want to find a policy π^* for which

$$v_N^{\pi^*}(s) \geq v_N^\pi(s) \quad \text{for} \quad s \in S \quad \text{and} \quad \pi \in \Pi^{HR}.$$

Such a policy is called a **optimal policy**. If this optimal policy π^* does not exist, then we want to find an $\epsilon-$**optimal policy**. An ϵ-optimal policy π_ϵ^* is a policy with the following property

$$v_N^{\pi_\epsilon^*}(s) + \epsilon > v_N^\pi(s) \quad \text{for} \quad \epsilon > 0, \quad s \in S \quad \text{and} \quad \pi \in \Pi^{HR}.$$

Further, for an optimal policy π^*, the **value** of the Markov decision problem v_N^* is defined as

$$v_N^*(s) \coloneqq \sup_{\pi \in \Pi^{HR}} v_N^\pi(s) \quad \text{for} \quad s \in S. \tag{3.19}$$

If we do not have a optimal policy, but only an ϵ-optimal policy π_ϵ^*, then the value is given by

$$v_N^{\pi_\epsilon^*}(s) + \epsilon > v_N^*(s) \quad \text{for} \quad \epsilon > 0 \quad \text{and} \quad s \in S. \tag{3.20}$$

For an an optimal policy π^*, the **expected total reward** is given by

$$v_N^{\pi^*}(s) = v_N^*(s) \quad \text{for} \quad s \in S. \tag{3.21}$$

Lastly, we will discuss optimality equations, which are also known as the Bellman equations. For a history dependent and randomized policy π we define the **optimal value functions** $u_t^* : H_t \to \mathbb{R}$ by

$$u_t^* \coloneqq \sup_{\pi \in \Pi^{HR}} u_t^\pi(h_t)$$

with $u_t^{\pi}(h_t)$ the expected total reward. The **optimality equations** or **Bellman equations** are then given by

$$u_t(h_t) = \sup_{a \in A_{s_t}} \{r_t(s_t,a) + \sum_{j \in S} p_t(j \mid s_t,a) u_{t+1}(h_t,a,j)\} \text{ for } t = 1,...,N-1 \text{ and } h_t = (h_{t-1},a_{t-1},s_t) \in H_t. \tag{3.22}$$

For $t = N$, we have

$$u_N(h_N) = r_N(s_N) \text{ for } h_N = (h_{N-1},a_{N-1},s_N) \in H_N, \tag{3.23}$$

which we refer to as the boundary condition. The importance of the Bellman equations is that they are used to verify that a given policy is optimal. Now, we will state a simple lemma, which will later be used in an important theorem.

Lemma 3.2.1. Let w be a real-valued function on an arbitrary discrete set W and suppose that $q(\cdot)$ is a probability distribution on that arbitrary discrete set W. Then

$$\sup_{u \in W} w(u) \geq \sum_{u \in W} q(u) w(u).$$

Proof:
Define $w^* = \sup_{u \in W} w(u)$. Then,

$$w^* = \sum_{u \in W} q(u) w^* \geq \sum_{u \in W} q(u) w(u).$$

■

Theorem 3.2.1. Suppose u_t, $u_t : H_t \to R$, is a solution of the Bellman equations stated in (3.22) for $t = 1,...N-1$ and suppose that u_N satisfies the boundary condition (3.23). Then

a. $u_t(h_t) = u_t^*(h_t) \ \ \forall h_t \in H_t$ and $t = 1,...,N$

b. $u_1(s_1) = v_N^*(s_1) \ \ \forall s_1 \in S$

Part (a) implies that the solutions of the Bellman equations are the optimal value functions from time t onward and part (b) implies that the solution obtain at time $t = 1$ is the value of the Markov decision problem.

Proof:
We start with proving part (a), which will be done by proving the following two claims.

Claim 1: $u_t(h_t) \geq u_t^*(h_t) \ \ \forall h_t \in H_t$ and $t = 1,...,N$

Claim 2: $\forall \epsilon > 0 \ \ \exists \pi'' \in \Pi^{HD}$ (π'' is a history dependent and deterministic policy) for which

$$u_t^{\pi''}(h_t) + (N-t)\epsilon \geq u_t(h_t) \ \ \forall h_t \in H_t \text{ and } t = 1,...,N$$

We will now proof both claims by backwards induction on t.
Proof of claim 1:
For the induction start, we observe that for $t = N$ the result holds true, because there are no decisions made in period N, hence we have the boundary condition

$$u_N(h_N) = r_N(s_N) = u_N^{\pi}(h_N) \text{ for } h_N = (h_{N-1},a_{N-1},s_N) \in H_N \text{ and } \pi \in \Pi^{HR}.$$

Therefore,

$$u_N(h_N) = u_N^*(h_N) \ \ \forall h_N \in H_N.$$

For the induction hypothesis, we assume that the result holds true for $t = n+1, ..., N$, hence we assume $u_t(h_t) \geq u_t^*(h_t) \quad \forall h_t \in H_t$ and $t = n+1, ..., N$.
For the induction step, we will prove that the result holds true for $t = n$. Suppose that $\pi' = (d_1', d_2', ..., d_{N-1}')$ is an arbitrary policy in Π^{HR}. Then, the Bellman equation for $t = n$ is given by

$$u_n(h_n) = \sup_{a \in A_{s_n}} \{r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) u_{n+1}(h_n, a, j)\}.$$

Hence, by applying the induction hypothesis and lemma (3.2.1), we obtain the following:

$$\begin{aligned} u_n(h_n) &= \sup_{a \in A_{s_n}} \{r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) u_{n+1}(h_n, a, j)\} \\ &\geq \sup_{a \in A_{s_n}} \{r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) u_{n+1}^*(h_n, a, j)\} \\ &\geq \sup_{a \in A_{s_n}} \{r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) u_{n+1}^{\pi'}(h_n, a, j)\} \\ &\geq \sum_{a \in A_{s_n}} q_{d_n'(h_n)}(a)(r_n(s_n, a) + \sum_{j \in S} p_n(j \mid s_n, a) u_{n+1}^{\pi'}(h_n, a, j)) \\ &= u_n^{\pi'}(h_n) \end{aligned}$$

Since the policy $\pi' = (d_1', d_2', ..., d_{N-1}')$ is an arbitrary policy in Π^{HR}, we have

$$u_n(h_n) \geq \sup_{\pi \in \Pi^{HR}} u_n^{\pi}(h_n) = u_n^*(h_n).$$

The last equality is by definition of the optimal value functions. So we have proven claim 1.

⊠

Proof of claim 2:
Let $\pi'' = (d_1'', d_2'', ..., d_{N-1}'')$ be a policy that is constructed by choosing $d_t''(h_t) = a$ for $a \in A_{s_t}$, such that

$$r_t(s_t, d_t''(h_t)) + \sum_{j \in S} p_t(j \mid s_t, d_t''(h_t)) u_{t+1}^{\pi''}(h_t, d_t''(h_t), j) + \epsilon \geq u_t(h_t).$$

This is possible, if we assume that $u_t(h_t)$ satisfies the Bellman equations.
This claim will also be proven by induction.
For the induction start, we observe that the result holds true for $t = N$, since $u_N^{\pi''}(h_N) = u_N(h_N)$.
For the induction hypothesis, we assume that the result holds true for $t = n+1, ..., N$, hence we assume $u_t^{\pi''}(h_t) + (N-t)\epsilon \geq u_t(h_t)$ for $t = n+1, ..., N$.
For the induction step, we will prove that the result holds true for $t = n$. So

$$\begin{aligned} u_n^{\pi''}(h_n) &= r_n(s_n, d_n''(h_n)) + \sum_{j \in S} p_n(j \mid s_n, d_n''(h_n)) u_{n+1}^{\pi''}(h_n, d_n''(h_n), j) \\ &\geq r_n(s_n, d_n''(h_n)) + \sum_{j \in S} p_n(j \mid s_n, d_n''(h_n)) u_{n+1}(h_n, d_n''(h_n), j) - (N-n-1)\epsilon \\ &\geq u_n(h_n) - (N-n)\epsilon. \end{aligned}$$

Hence,

$$u_n^{\pi''}(h_n) + (N-n)\epsilon \geq u_n(h_n).$$

So we have proven claim 2.

⊠

Claims 1 and 2 together show that for any $\epsilon > 0$, there exists a policy $\pi \in \Pi^{HR}$ such that

$$u_t^*(h_t) + (N-t)\epsilon \geq u_t^{\pi}(h_t) + (N-t)\epsilon \geq u_t^{\pi}(h_t) \geq u_t^*(h_t)$$

for $t = 1, ..., N$ and $h_t \in H_t$. Since N is a fixed number, we let ϵ approach zero, hence

$$u_t(h_t) = u_t^*(h_t).$$

Part (b) follows from the fact that $u_1(s_1) = u_1^*(s_1) = v_N^*(s_1)$.

■

In the following section, we will discuss the infinite-horizon model.

3.3 Infinite-Horizon Markov Decision Processes

In this section, we will explain discrete-time infinite-horizon Markov decision processes, hence $T = \{1, 2, ...\}$. We will introduce different criteria, such as the expected total reward criterion, the expected total discounted reward criterion and optimality criteria. Throughout this discussion we will assume that our data is time-homogeneous, which means that we assume that the rewards $r_t(s,a)$, the transition probabilities $p_t(j \mid s,a)$ and the set of decision rules D_t^K ($K \in \{MD, HD, MR, HR\}$) do not change over time. Hence, for all $t \in T$, $r_t(s,a) = r(s,a)$, $p_t(j \mid s,a) = p(j \mid s,a)$ and $\pi = d^{\infty} = (d, d, ...)$. The latter means that in every decision epoch we have the same decision rule. Before, we start discussing the criteria, we will discuss the **value of a policy**. Given a discrete time-homogeneous infinite-horizon Markov decision processes. Each policy $\pi = (d_1, d_2, ...)$ induces a **reward process**, denoted by $\{(X_t, r_t(X_t, Y_t)) \mid t \in T\}$ as we have seen in section one of chapter three. The first component of the reward process, X_t, represents the state of the process at time t. The second component, $r_t(X_t, Y_t)$, represents the reward received in state X_t at time t, when action Y_t is used. The action Y_t is determined by the decision rule d_t in the following way:

- For $d_t \in D^{MD}$, a Markovian and deterministic decision set, we have

$$Y_t = d_t(X_t).$$

- For $d_t \in D^{HD}$, a history-dependent and deterministic decision set, we have

$$Y_t = d_t(Z_t).$$

  The random variable Z_t denotes the history up to time t as we have seen before.

- For $d_t \in D^{MR}$, a Markovian and randomized decision set, we have

$$\mathbb{P}(\{Y_t = a\}) = q_{d_t(X_t)}(a).$$

- For $d_t \in D^{HR}$, a history-dependent and randomized decision set, we have

$$\mathbb{P}(\{Y_t = a\}) = q_{d_t(Z_t)}(a).$$

We will now discuss how to assign a value to a policy $\pi \in \Pi^{HR}$, which we also did for the discrete-time finite-horizon Markov decision process. The **expected total reward** for a history dependent

and randomized policy π with state s as initial state is defined as

$$v^{\pi}(s) \coloneqq \lim_{N\to\infty} \mathbb{E}_s^{\pi}[\sum_{t=1}^{N} r_t(X_t, Y_t)] = \lim_{N\to\infty} v_{N+1}^{\pi}(s). \tag{3.24}$$

Note that this limit does not need to exist and may be $\pm\infty$. If the limit exist and we may interchange limit and expectation, then we may write

$$v^{\pi}(s) = \mathbb{E}_s^{\pi}[\sum_{t=1}^{\infty} r_t(X_t, Y_t)]. \tag{3.25}$$

For a history dependent and randomized policy π, the **expected total discounted reward** is defined by

$$v_{\lambda}^{\pi}(s) \coloneqq \lim_{N\to\infty} \mathbb{E}_s^{\pi}[\sum_{t=1}^{N} \lambda^{t-1} r_t(X_t, Y_t)] \tag{3.26}$$

for $0 \leq \lambda < 1$ the discount factor.
Note that this limit only exists when $\sup_{s\in S}\sup_{a\in A_S}|r(s,a)| \leq \infty$.
If the limit exists and we may interchange limit and expectation, then we may write

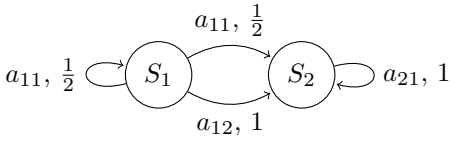
$$v_{\lambda}^{\pi}(s) = \mathbb{E}_s^{\pi}[\sum_{t=1}^{\infty} \lambda^{t-1} r_t(X_t, Y_t)]. \tag{3.27}$$

Lastly, the **average reward** of a history dependent and randomized policy π is defined by

$$g^{\pi}(s) \coloneqq \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_s^{\pi}[\sum_{t=1}^{N} r_t(X_t, Y_t)] = \lim_{N\to\infty} \frac{1}{N} v_{N+1}^{\pi}(s). \tag{3.28}$$

We will now consider the following example.

Example 3.3.1. This is a sequel of example (3.1.2). Again, consider the following representation of the two state Markov decision process.



In this example, we assume that the rewards and transition probabilities are the same at each epoch. There are two states $S = \{S_1, S_2\}$. In state S_1, the decision maker chooses either action a_{11} or action a_{12}. In state S_2 the only choice the decision maker has is action a_{21}. Choosing action a_{11} in state S_1 leads to an immediate reward of five units, and the system will evolve to state S_1 with a probability of $\frac{1}{2}$ and to state S_2 with a probability of $\frac{1}{2}$ as well. If the decision maker chooses action a_{12} in state S_1, then he will receive an immediate reward of ten units and the system evolves to state S_2 with a probability of 1. In state S_2, the decision maker has no other choice than to choose action a_{21}, by doing so he will receive an immediate reward of minus one unit and the system stays in state S_2 with probability 1.
There are two deterministic Markovian decision rules, namely d_1 and d_2.
For state S_1, we have $d_1^{MD}(s_1) = a_{11}$ and $d_2^{MD}(s_1) = a_{12}$.
And for state S_2, we have $d_1^{MD}(s_2) = a_{21}$ and $d_2^{MD}(s_2) = a_{21}$.
The rewards for $N \geq 1$ are given by $v_N^{d_1^{\infty}}(s_1) = 5 - 0.5 \cdot (N-1)$, $v_N^{d_2^{\infty}}(s_1) = 10 - 1 \cdot (N-1)$ and $v_N^{d_1^{\infty}}(s_2) = v_N^{d_2^{\infty}}(s_2) = -1 - 1 \cdot (N-1) = -N$.

Hence, the expected total reward is equal to

$$\lim_{N\to\infty} v_N^{d_1^\infty}(s_1) = \lim_{N\to\infty} v_N^{d_2^\infty}(s_1) = \lim_{N\to\infty} v_N^{d_1^\infty}(s_2) = \lim_{N\to\infty} v_N^{d_2^\infty}(s_2) = -\infty.$$

For the expected total discounted reward, it can be shown that

$$v_\lambda^{d_1^\infty}(s_1) = \frac{5-5.5\lambda}{(1-0.5\lambda)(1-\lambda)}, \quad v_\lambda^{d_1^\infty}(s_2) = -\frac{1}{1-\lambda}$$

$$v_\lambda^{d_2^\infty}(s_1) = 10 - \frac{\lambda}{1-\lambda}, \quad v_\lambda^{d_2^\infty}(s_2) = -\frac{1}{1-\lambda}.$$

We will collaborate on $v_\lambda^{d_2^\infty}(s_2) = -\frac{1}{1-\lambda}$. We know that $\sum_{t=1}^{N} \lambda^{t-1} = \frac{\lambda^N - 1}{\lambda - 1}$. We also know that $v_N^{d_2^\infty}(s_2) = -N$. Therefore,

$$\lim_{N\to\infty} \sum_{t=1}^{N} \lambda^{t-1} \cdot -N = \lim_{N\to\infty} \frac{\lambda^N - 1}{\lambda - 1} \cdot -N = \frac{1}{\lambda - 1} = -\frac{1}{1-\lambda}.$$

The average reward of both deterministic Markovian policies, d_1 and d_2, is equal to -1, since we have an absorption in state S_2, hence $g^{d_1^\infty}(s_1) = g^{d_2^\infty}(s_1) = g^{d_1^\infty}(s_2) = g^{d_2^\infty}(s_2) = -1$.

♦

We will now start with the **expected total reward criterion**. The goal is to find a policy π with the largest value of

$$v^\pi(s) = \lim_{N\to\infty} v_N^\pi(s). \tag{3.29}$$

The only problem is that this limit does not need to exist, as we have seen before. We will solve this problem by providing certain conditions for rewards $r(s,a)$ and transition probabilities $p(j \mid s,a)$, such that they ensure the existence of the limit (3.29). We will define the following quantities:

$$v_+^\pi(s) \coloneqq \mathbb{E}_s^\pi[\sum_{t=1}^{\infty} r^+(X_t, Y_t)] \tag{3.30}$$

and

$$v_-^\pi(s) \coloneqq \mathbb{E}_s^\pi[\sum_{t=1}^{\infty} r^-(X_t, Y_t)], \tag{3.31}$$

with $r^+(s,a) \coloneqq \max\{+r(s,a), 0\}$ and $r^-(s,a) \coloneqq \max\{-r(s,a), 0\}$.
The values of $r^+(s,a)$ as well as $r^-(s,a)$ are non-negative, hence both quantities $v_+^\pi(s)$ and $v_-^\pi(s)$ are guaranteed to exists, but could be equal to infinity. To rule out the possibility that both quantities $v_+^\pi(s)$ and $v_-^\pi(s)$ are infinite, we need to make the following assumption. For all history dependent and randomized policies π and for all $s \in S$, at least one of the quantities $v_+^\pi(s)$ and $v_-^\pi(s)$ is finite. Under this assumption, we have that the limit (3.29) exists and that the expected total reward is equal to $v^\pi(s) = v_+^\pi(s) - v_-^\pi(s)$.
We will now formulate a definition, in which we categorize the several classes of models that satisfies the given assumption.

Definition 3.3.1. Suppose we have a discrete time-homogeneous infinite-horizon Markov decision process. Then we can categorize the following classes of models:

- A process belongs to the class of **positive bounded models** if for each $s \in S$, there exists an $a \in A_S$, such that $r(s,a) \geq 0$ and $v_+^\pi(s)$ is finite for all history dependent and randomized policies π
- A process belongs to the class of **negative models** if for each $s \in S$ and for all $a \in A_S$, the reward $r(s,a) \leq 0$ and $v^\pi(s) > -\infty$ for some history dependent and randomized policy π
- A process belongs to the class of **convergent models** if for each $s \in S$, the quantities of $v_+^\pi(s)$ as well as $v_-^\pi(s)$ are finite for all history dependent and randomized policies π

The limit (3.29) exists for a process that belongs to one of these classes. A positive bounded model has the property that there exists a stationary policy with a non-negative expected total reward, so as a consequence the optimal value function will be non-negative. For a negative model we have that $v_+^\pi(s) = 0$ for all history dependent and randomized policy π. So for a negative model, our goal is to find a policy π that minimizes $v_-^\pi(s)$. And for a convergent model, we require that both quantities $v_+^\pi(s)$ and $v_-^\pi(s)$ are finite. Which means that

$$v^\pi(s) = \mathbb{E}_s^\pi[\sum_{t=1}^{\infty} |r(X_t, Y_t)|] = v_+^\pi(s) - v_-^\pi(s) < \infty$$

holds for all history dependent and randomized policies π and states $s \in S$.
We will now discuss the **expected total discounted reward criterion**. The discount factor $\lambda \in [0,1)$ is a measure for the present value of one unit with respect to the future value of that unit. So if we have 10 units in the present state and our discount factor is $\lambda = 0.5$, then in the next state we have $0.5 \cdot 10 = 5$ units. The value $v_\lambda^\pi(s)$ denoted the expected total discounted reward as we have seen before. We will now define another value in which we make use of the horizon length μ, that follows a geometric distribution with parameter $\lambda \in [0,1)$ and is independent of the policy π. This geometric distribution is given by $\mathbb{P}(\mu = n) = (1-\lambda)\lambda^{n-1}$ for $n = 1, 2, ...$
Let $v_\mu^\pi(s)$ denote the expected total reward that is obtained by using policy π when the horizon length μ is random and independent of the chosen actions. Then $v_\mu^\pi(s)$ is defined by:

$$v_\mu^\pi(s) \coloneqq \mathbb{E}_s^\pi[\mathbb{E}_\mu[\sum_{t=1}^{\mu} r(X_t, Y_t)]]. \tag{3.32}$$

Proposition 3.3.1. Suppose that the limit of $v_\lambda^\pi(s)$ (3.26) exists and suppose that μ has a geometric distribution with parameter λ. Then

$$v_\lambda^\pi(s) = v_\mu^\pi(s) \quad \text{for all} \quad s \in S.$$

Proof:
We know that $v_\lambda^\pi(s) = \mathbb{E}_s^\pi[\sum_{t=1}^{\infty} \lambda^{t-1} r_t(X_t, Y_t)]$ (3.27) if we may interchange limit and expectation. Further, $v_\mu^\pi(s) \coloneqq \mathbb{E}_s^\pi[\mathbb{E}_\mu[\sum_{t=1}^{\mu} r(X_t, Y_t)]]$ (3.32).
We will rewrite $v_\mu^\pi(s)$ as follows:

$$v_\mu^\pi(s) \coloneqq \mathbb{E}_s^\pi[\mathbb{E}_\mu[\sum_{t=1}^{\mu} r(X_t, Y_t)]] =$$

$$\mathbb{E}_s^\pi[\sum_{n=1}^{\infty} (1-\lambda)\lambda^{n-1} \sum_{t=1}^{n} r(X_t, Y_t)] =$$

$$\mathbb{E}_s^\pi[\sum_{t=1}^{\infty} r(X_t, Y_t) \sum_{n=t}^{\infty} (1-\lambda)\lambda^{n-1}] =$$

$$\mathbb{E}_s^\pi[\sum_{t=1}^{\infty} r(X_t, Y_t) \cdot (1-\lambda) \sum_{n=t}^{\infty} \lambda^{n-1}] =$$

Note that $\sum_{n=1}^{\infty} \lambda^{n-1} = \frac{1}{1-\lambda}$, hence $\sum_{n=t}^{\infty} \lambda^{n-1} = \frac{\lambda^{t-1}}{1-\lambda}$.
So

$$\upsilon_\mu^\pi(s) = \mathbb{E}_s^\pi[\sum_{t=1}^{\infty} r(X_t, Y_t) \cdot (1-\lambda) \sum_{n=t}^{\infty} \lambda^{n-1}] =$$

$$\mathbb{E}_s^\pi[\sum_{t=1}^{\infty} r(X_t, Y_t) \cdot (1-\lambda) \cdot \frac{\lambda^{t-1}}{1-\lambda}] =$$

$$\mathbb{E}_s^\pi[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)] = \upsilon_\lambda^\pi(s)$$

■

The last subject we will discuss in this section is optimality criteria. As we have seen before, a policy is called a optimal policy if the value function of that policy is the largest. We will discuss those optimality criteria in the following definition.

Definition 3.3.2. Suppose we have a discrete time-homogeneous infinite-horizon Markov decision process. We assume that π^* is an history-dependent and randomized policy.

- The value of the Markov decision process is defined as

  $$\upsilon^*(s) \coloneqq \sup_{\pi \in \Pi^{HR}} \upsilon^\pi(s) \quad \text{for} \quad s \in S$$

  if the limit of expected total reward $\upsilon^\pi(s)$ exist, see (3.19) and (3.25). We call a policy π^* a **total reward optimal** policy if

  $$\upsilon^{\pi^*}(s) \geq \upsilon^\pi(s) \quad \text{for all} \quad s \in S \quad \text{and} \quad \pi \in \Pi^{HR}.$$

- The discounted value of the Markov decision process is defined as

  $$\upsilon_\lambda^*(s) \coloneqq \sup_{\pi \in \Pi^{HR}} \upsilon_\lambda^\pi(s) \quad \text{for} \quad s \in S$$

  if the limit of expected total discounted reward $\upsilon_\lambda^\pi(s)$ exist, see (3.26). We call a policy π^* a **discount optimal** policy if

  $$\upsilon_\lambda^{\pi^*}(s) \geq \upsilon_\lambda^\pi(s) \quad \text{for all} \quad s \in S \quad \text{and} \quad \pi \in \Pi^{HR}.$$

- The optimal gain of the Markov decision process is defined as

  $$g^*(s) \coloneqq \sup_{\pi \in \Pi^{HR}} \upsilon^\pi(s) \quad \text{for} \quad s \in S$$

  if the limit of the gain $g^\pi(s)$ exist, see (3.28). We call a policy π^* a **gain optimal** policy if

  $$g^{\pi^*}(s) \geq g^\pi(s) \quad \text{for all} \quad s \in S \quad \text{and} \quad \pi \in \Pi^{HR}.$$

In the next chapter, we will discuss partially observed Markov decision processes.

Chapter 4

Partially Observed Markov Decision Processes

In this chapter, we will introduce the basic components of a discrete-time partially observed Markov decision process. A partially observed Markov decision process differs from a Markov decision process in the way that not all information is available to the decision maker. The decision maker is unsure in which state he is, he only receives an observation and a reward. The decision maker bases his decisions on observations and past actions.

A partially observed Markov decision process consists of seven elements: decision epochs or times, states, actions, rewards, transition probabilities, observations and observation probabilities. The first five components we have seen in the model formulation of Markov decision processes in section one of chapter three. Besides the decision epochs, states, actions, rewards, transition probabilities, observations and observation probabilities we have a decision maker that observes the process and may select actions at each decision epoch to influence the system and gain rewards. The only differences here is that not all information is available as we said before. So mathematically, we can formulate a partially observed Markov decision process by the collection of the seven elements

$$\{T, S, A_s, O, p_t(\cdot \mid s,a), p_o(\cdot \mid s), r_t(s,a) \mid t \in T, s \in S, a \in A_s, o \in O\}.$$

The seven elements and its notation will be explained below, the first five components will be discussed briefly since we have seen it before. After that we will discuss an example of the difference between a Markov decision process and a partially observed Markov decision process. We will now start with the explanation of the components of a discrete-time partially observed Markov decision process.

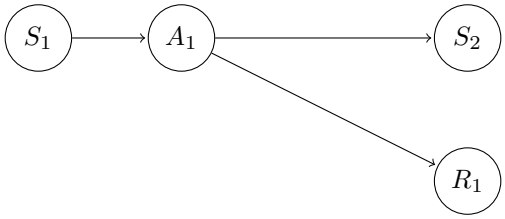
As we have seen before, **Decision epochs** or **decision times** are given points in time, where decisions are made by the decision maker. We denote T as the set of decision times and we assume that the set T is discrete. We make this assumption, because we are only interested in discrete-time partially observed Markov decision processes. Further, we denote the elements of the set T as t, which we refer to as time t. Secondly, we will explain the state, the observation and actions set. The process occupies a **state** at each decision time, we denote the set of all possible states by S. The elements of S will be denoted by s.The difference between a Markov decision process and a partially observed Markov decision process is that the states will not be observed, the decision maker does not know in which state he is. So instead of a state, the decision maker receives an **observation** at each decision epoch, we denote the set of all observations by O. The elements of the set of observations will be denoted by o. The probability that the decision make observes a certain observation is given by the conditional **observation probability** $p_O(o \mid s,a)$. When the

decision maker observes the system at a certain observation $o \in O$ at a given decision time t, then he may choose an **action** $a \in A_s$, where A_s is the set of possible actions in state s even though he did not observed that state. Lastly, we will discuss the rewards and the transition probabilities. The **reward function** $r_t(s,a)$ denotes the value of the reward received at time $t \in T$ for $s \in S$ and $a \in A$. The **transition probability function** $p_t(j \mid s,a)$ denotes the probability that the system will be in state j at the next decision epoch $t+1$, when the decision maker chooses an action $a \in A_s$ in state $s \in S$ at decision time t.
We have discussed the components of a discrete-time partially observed Markov decision process model. So we can conclude that the partially observed Markov decision process can be formulated by the collection of the seven elements, hence

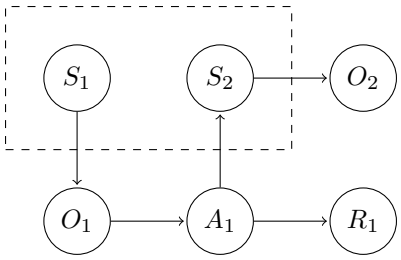
$$\{T, S, A_s, O, p_t(\cdot \mid s,a), p_o(\cdot \mid s,a), r_t(s,a) \mid t \in T, s \in S, a \in A_s, o \in O\}.$$

We will now discuss a really simplistic example that elaborates the difference between a Markov decision process and a partially observed Markov decision process.

Example 4.0.1. In this example, we will give a simple representation of a Markov decision process and a partially observed Markov decision process. Each figure consists of two states S_1 and S_2 and one action A_1 and one reward R_1.
We will start with Markov decision process.



In this figure, we see that the decision maker observes state S_1 and chooses action A_1, which gives him a reward R_1 and the system will evolve to the next state S_2.
Now, we will give a representation of a partially observed Markov decision process. Please note that within the dashed box is not observed by the decision maker.



In this figure, we see that the decision maker receives observation O_1 and chooses action A_1, which gives him a reward R_1 and the system will evolve to the next state S_2. The decision maker does not observe state S_2, but again he receives an observation, O_2.
So the difference between the two is that in the second case the states are not observed.

♦

In the next chapter, we will discuss the applications of Markov decision processes in the medical sciences.

Chapter 5

Applications in Medical Sciences

In this chapter, we will discuss the applications of Markov decision processes in the medical sciences. We will discuss two different cases, namely infectious diseases and ischemic heart disease.

5.1 Infectious Diseases

In this section, we will discuss the SIR-model and Markov decision process for an influenza epidemic. We will start with introducing some terminology. In order to control the spread of an emerging infectious disease, such as influenza, we need **health policies**. A health policy makes real-time recommendations, in order to respond to changing disease characteristics, population characteristics and resource constraints. One could think of infectivity and resistance to antibiotics as disease characteristics. For population characteristics one could imagine disease prevalence and the proportion of individuals that are immune for the disease. Lastly, vaccines, antibiotics, budget and health care staff are examples of resource constraints. These health policies allow the decision maker to use the current data from the epidemic and the resource availability to make decisions or interventions at any point in time. The goal is to determine the optimal health policy for controlling the spread of infectious disease during an epidemic. There are several things that may be involved by the control of an influenza epidemic, namely reducing susceptibility of uninfected individuals, reducing contact rates in the population and reducing the infectiousness of infected individuals. This may be done through vaccination, isolation and treatment. Further, the control of an emerging influenza epidemic is bounded by the availability of vaccines and the availability of money and resources for vaccine procurement, diagnosis and treatment of new cases. With those resource constraints in our minds, we define the optimality of a health policy as the efficient use of available resources to maximize the overall health of the population.

We start with describing the influenza epidemic by using a SIR model, this model will be reformulated as a Markov decision process further on in this section. A SIR model or a Susceptible-Invective-Recovered model assumes that individuals who are recovered from the infection are now permanent immune to that specific infection. The SIR model consists of the five components, namely states, decision sets, actions, rewards and transition probabilities. After we have discuss these five components, we will discuss decision rules, health policies and optimally.

Again, we will start with some terminology. **Decision times** are given points in time, where decisions are made by the decision maker. We denote T as the set of decision times and we assume that the set T is discrete. The set of decision epochs T can either be finite of infinite, hence $T = \{1, 2, ..., M\}$ or $T = \{1, 2, ...,\}$ respectively. In the first case, we have a finite horizon and in the second case an infinite horizon. We have seen this before in chapter three section one.

We denote $X_S(t)$ as the number of susceptibles at time t, $X_I(t)$ as the number of infectives at time t and $X_R(t)$ as the number of recovered individuals at time t. Further, we will make the assumption that the population size does not change during the epidemic. In this case, we denote the population size by N.

Now, we will discuss disease states and actions. The **disease state** or the state of the disease spread is denoted by $s_t = (X_S(t), X_I(t))$ for any given time t. The set of all states or **state space** is denoted by S, which is defined as $S \coloneqq \{(x_S, x_I) \in \mathbb{N}^2 \mid x_S + x_I \leq N\}$. There are two possible interventions or **actions** to control the spread of the infectious disease, namely vaccination and transmission-reducing intervention. We can implement vaccination in our model as follows. The decision maker has the opportunity to select a number of susceptibles to vaccinate at any decision time conditional on the availability and the price of the vaccine. The decision to immunize is denoted by $z_t \in A_I$, with $A_I = [0, N]$ the set of all possible numbers of susceptibles to vaccinate. We assume that vaccination at decision time t result in immunization at decision time $t+1$. So the number of susceptibles at decision time t is $X_S(t)$ and at the next decision epoch the number reduces to $X_S(t+1) = X_S(t) - z_t$. Further, the implementation of transmission-reducing interventions can be done as follows. Transmission-reducing interventions may include social distancing, hygienic interventions and treatment or isolation. The set of transmission-reducing interventions is denoted by $A_T = \{0, 1, ..., M\}$, where $A_T = 0$ implies no interventions at all. The decision to employ transmission-reducing interventions is denoted as $a_t \in A_T$ at decision time t. As we have seen before, the number of susceptibles $X_S(t)$ at decision time t reduces to $X_S(t+1) = X_S(t) - z_t$ at decision time $t+1$ when the decision maker chooses to employ vaccination at time t, $t = \{0, 1, 2, ...\}$. We have now discussed states and actions. In the next paragraph, we will elaborate on rewards and transition probabilities.

When the decision maker chooses to employ vaccination at time t, he receives a **reward** $r_t(s_t, z_t)$, $r_t(s_t, z_t) = -pz_t$, where p is the unit price of a vaccine and z_t the number of susceptibles that are immunized at time t. If the decision maker chooses to employ transmission-reducing interventions, $a_t \in A_T$, at decision time t, then the disease spread at decision time $t+1$ is determined by the **transition probability** $p_t(\cdot \mid s_t, a_t)$. In this case, the decision maker receives a **reward** $r_t(s_t, a_t)$. To define this reward, we need to define some additional parameters, since the control of an epidemic may be bounded by the availability of medical and monetary resources. These additional parameters are:

- λ: The willingness to pay for health. This is a constant.
- c: The cost incurred for each infection. This is a constant.
- $c_T(a_t)$: The cost for implementation of transmission-reducing interventions $a_t \in A_T$ at decision time t. We assume that $c_T(0) = 0$, hence if the decision maker does not intervene then there are no cost.
- $u(s_t, a_t)$: The expected cost incurred during period t if the disease spread at time t is at state s_t and the decision maker chooses to implement the transmission-reducing interventions $a_t \in A_T$ at decision epoch t.
- w: The loss in health due to infections. This is also a constant.
- $l(s_t, a_t)$: The expected loss in health of the population during period t if the disease spread at time t is at state s_t, and the decision maker chooses to implement transmission-reducing interventions $a_t \in A_T$ at decision epoch t. So $l(s_t, a_t) = w\mathbb{E}[I(t) \mid s_t, a_t]$ with $I(t)$ the number of new infections during period t.

We can now define the reward $r_t(s_t, a_t)$ received by the decision maker as
$r_t(s_t, a_t) = \lambda l(s_t, a_t) - u(s_t, a_t)$. Now we have discussed the five components of the SIR-model, we will continue with decision rules, health policies and optimality.

As we have seen in chapter three section one, a **decision rule** prescribes an action for each state for a given decision time. We will now describe the decision rules. For decision time $t = 0$, the decision rule is defined as a function $d_0 : S \to A_I$. This function specifies the number of susceptibles $z_0 \in A_I$ to vaccinate given the initial disease state $s_0 = (X_S(0), X_I(0)) \in S$. For decision time $t = \{1, 2, ...\}$, the decision rule is defined as a function $d_t : S \to A_T \times A_I$. This function specifies for each disease state a transmission-reducing intervention $a_t \in A_T$ and a number of susceptibles $z_t \in A_I$ to vaccinate. This decision rule is Markovian, since it depends only on the current state of disease spread. Throughout this discussion we will assume that the decision rules are Markovian.

Now, we will discuss health policies. A **health policy** π is a sequence of decision rules, so $\pi = (d_0, d_1, ...)$. We denote Π as the set of all health policies. Further, a health policy is called **stationary** if $d_t = d$ for all $t \in T$. This means that a health policy prescribes the same decision d at every decision time t regardless of the state s_t. Throughout this discussion we assume that all health policies are stationary.

If we assume that the influenza spread is at state s_1 at decision epoch $t = 1$, then the **expected total discounted reward** is defined as

$$\upsilon_\gamma^\pi(s_1) \coloneqq \mathbb{E}_{s_1}^\pi[\sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t, d_t(s_t)) \mid s_1] \quad \text{with} \quad \gamma \in [0,1] \quad \text{the discount factor.}$$

We have seen something similar, when we discussed infinite-horizon Markov decision processes in section three of chapter three, (3.26). If we now assume that the influenza spread is at state s_0 at decision epoch $t = 0$, then the **expected total reward** is defined as

$$\Upsilon^\pi(s_0) \coloneqq r_t(s_0, d_0(s_0)) + \gamma \upsilon^\pi(s_1) \quad \text{with} \quad \gamma \in [0,1] \quad \text{the discount factor.}$$

Again, we saw something similar before in section three of chapter three, (3.24).
A health policy π is said to be an **optimal health policy** π^* if

$$\Upsilon^{\pi^*}(s_0) \geq \Upsilon^\pi(s_0) \quad \text{for all} \quad s_0 \in S \quad \text{and} \quad \pi \in \Pi.$$

The model we just described is called a SIR-model, with some modifications we can use a Markov decision process to find the optimal health policy π^*. If we can use a discrete-time Markov chain to model the disease dynamics and if the states are observable throughout the epidemic, then we can obtain the optimal health policy π^* by using a Markov decision process. To be able to use a Markov decision process for describing an influenza epidemic, we need to make some simplifying assumptions. First, we assume that the population size does not change during the epidemic. Secondly, we assume that individuals only become infected through contact with other infected individuals. Thirdly, we assume that contacts occur according to a homogeneous Poisson distribution with rate $\mu\Delta t$ for $t \in T$ during the time interval $[t, t + \Delta t]$. Lastly, we assume that a susceptible individual who is infected during time interval $[t - \Delta t, t]$ becomes infectious and symptomatic at time t. This individual will come in contact with other individuals during time interval $[t, t + \Delta t]$. The probability that a susceptible individual becomes infected when the individual comes in contact with an infected individual is denoted by $\alpha(t)$. Further, we denote the probability that the next interaction of a random susceptible individual is with an infected individual by $\beta(t)$. In the case no social distancing has occurred, $\beta(t)$ is equal to the number of the infected individuals $X_I(t)$ divided by the total population N, hence $\beta(t) = \frac{X_I(t)}{N}$. We can alter the variables $\alpha(t)$ and $\beta(t)$ by introducing transmission-reducing interventions. The overall probability that a susceptible individual becomes infected is denoted by $\varphi(t)$, with

$$\varphi(t) = 1 - e^{-\mu\Delta t\alpha(t)\beta(t)}. \tag{5.1}$$

Now, we will explain how we obtained this expression (5.1). We assume that a susceptible individual will come in contact with n individuals during time interval $[t, t+\Delta t]$. This contact will occur according to a homogeneous Poisson distribution with rate $\mu\Delta t$ for $t \in T$ as we have seen before. Further, we will assume that among those n individuals, j individuals will be infected. Lastly, the probability that the susceptible individual becomes infected is one minus the probability that none of the interactions with the j infected individuals results in infection, hence $1-(1-\alpha(t))^j$. Thus, we have a binomial distribution $(n, \beta(t))$. Therefore, $\varphi(t)$ is a composition of a Poisson distribution and a binomial distribution. So

$$\varphi(t) = \sum_{n=0}^{\infty} \frac{(\mu\Delta t)^n}{n!} \cdot e^{-\mu\Delta t} \left(\sum_{j=0}^{n} \binom{n}{j} (\beta(t))^j (1-\beta(t))^{n-j} (1-(1-\alpha(t))^j) \right) \tag{5.2}$$

Equation (5.2) can be rewritten as

$$\varphi(t) = 1 - \sum_{n=0}^{\infty} \frac{(\mu\Delta t)^n}{n!} \cdot e^{-\mu\Delta t} \left(\sum_{j=0}^{n} \binom{n}{j} (\beta(t))^j (1-\beta(t))^{n-j} (1-\alpha(t))^j \right) \tag{5.3}$$

We will now rewrite the expression $\sum_{j=0}^{n} \binom{n}{j} (\beta(t))^j (1-\beta(t))^{n-j} (1-\alpha(t))^j$ using the identity $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k}$. In our case, we choose x as $\beta(t)(1-\alpha(t))$ and y as $(1-\beta(t))$. Therefore,

$$\sum_{j=0}^{n} \binom{n}{j} (\beta(t))^j (1-\beta(t))^{n-j} (1-\alpha(t))^j =$$

$$(\beta(t)(1-\alpha(t)) + (1-\beta(t)))^n =$$

$$(\beta(t) - \alpha(t)\beta(t) + 1 - \beta(t))^n =$$

$$(1-\alpha(t)\beta(t))^n .$$

Now, we substitute this in our expression (5.3). Hence,

$$\varphi(t) = 1 - \left(\sum_{n=0}^{\infty} \frac{(\mu\Delta t)^n}{n!} \cdot e^{-\mu\Delta t} \right) (1-\alpha(t)\beta(t))^n \tag{5.4}$$

Lastly, we rewrite the expression $\sum_{n=0}^{\infty} \frac{(\mu\Delta t)^n}{n!} \cdot e^{-\mu\Delta t} \cdot (1-\alpha(t)\beta(t))^n$.

$$\sum_{n=0}^{\infty} \frac{(\mu\Delta t)^n}{n!} \cdot e^{-\mu\Delta t} \cdot (1-\alpha(t)\beta(t))^n =$$

$$e^{-\mu\Delta t} \cdot \sum_{n=0}^{\infty} \frac{(\mu\Delta t)^n (1-\alpha(t)\beta(t))^n}{n!} =$$

$$e^{-\mu\Delta t} \cdot \sum_{n=0}^{\infty} \frac{(\mu\Delta t(1-\alpha(t)\beta(t)))^n}{n!} =$$

$$e^{-\mu\Delta t} \cdot e^{(\mu\Delta t(1-\alpha(t)\beta(t))} =$$

$$e^{-\mu\Delta t + \mu\Delta t - \mu\Delta t\alpha(t)\beta(t)} =$$

$$e^{-\mu\Delta t\alpha(t)\beta(t)}.$$

Again, we will substitute this in our expression (5.4). Hence,

$$\varphi(t) = 1 - e^{-\mu\Delta t\alpha(t)\beta(t)}, \tag{5.5}$$

as required.

Given the disease state $s_t = (X_S(t), X_I(t))$, the number of new infections $I(t) = X_S(t) - X_S(t+\Delta t)$ during time interval $[t, t+\Delta t]$ will have a binomial distribution with $X_S(t)$ the number of trials and $\varphi(t)$ the probability of success. Thus,

$$\mathbb{P}(I(t) = k \mid X_S(t), X_I(t)) \coloneqq \begin{cases} \binom{X_S(t)}{k} (\varphi(t))^k (1-\varphi(t))^{X_S(t)-k} & \text{for } \ 0 \le k \le X_S(t), \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

We will now provide a framework to construct the transition probabilities of the Markov chain $\{(X_S(t), X_I(t)) \mid t = 0, 1, 2, ...\}$. The first step of the framework is to define the dynamics state equation. For a population of fixed size N, the dynamics state equation is defined to be $X_S(t) + X_I(t) + X_R(t) = N$, therefore two classes $X_S(t)$ and $X_I(t)$ are sufficient to construct a Markov model. The second step is to find the probability distribution of the driving events. We have two driving events in this Markov model, namely the number of new infections $I(t)$ during time interval $[t, t+\Delta t]$ and the number of recovered individuals $R(t)$ during time interval $[t, t+\Delta t]$. The probability distribution of the driving event $I(t)$ is given by equation (5.6). For the driving event $R(t)$, we assume that a susceptible individual who is infected during time interval $[t-\Delta t, t]$ becomes infectious and symptomatic at time t and will come in contact with other individuals during time interval $[t, t+\Delta t]$. At time $t+\Delta t$ this individual will be recovered and removed from the population. Hence, the probability distribution of the driving event $R(t)$ is defined as

$$\mathbb{P}(R(t) = r \mid X_S(t), X_I(t)) \coloneqq \begin{cases} 1 & \text{for } \ r = X_I(t), \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

The third step is to form the dynamics driving and feasibility constraints. The dynamics driving constraints will be defined as

$$I(t) = X_S(t) - X_S(t+\Delta t), \quad t \in T \tag{5.8}$$

$$R(t) = X_I(t) - X_I(t+\Delta t) + X_S(t) - X_S(t+\Delta t), \quad t \in T. \tag{5.9}$$

Further, the feasibility constraints are defined as follows

$$0 \le X_S(t) - X_S(t+\Delta t) \le X_S(t), \quad t \in T \tag{5.10}$$

$$0 \le X_I(t) - X_I(t+\Delta t) + X_S(t) - X_S(t+\Delta t) \le X_I(t), \quad t \in T. \tag{5.11}$$

The joint probability distribution of $(I(t), R(t))$, $\mathbb{P}(R(t), I(t) \mid X_S(t), X_I(t))$ is nonzero if and only if $R(t) = X_I(t)$. Since,

$$\mathbb{P}\big(R(t), I(t) \mid X_S(t), X_I(t)\big) =$$

$$\mathbb{P}(R(t), I(t) \mid X_S(t), X_I(t), R(t) = X_I(t)) \cdot \mathbb{P}(R(t) = X_I(t) \mid X_S(t), X_I(t)) +$$

$$\mathbb{P}(R(t), I(t) \mid X_S(t), X_I(t), R(t) \neq X_I(t)) \cdot \mathbb{P}(R(t) \neq X_I(t) \mid X_S(t), X_I(t)).$$

By equation (5.7) we know that

$$\mathbb{P}(R(t) = X_I(t) \mid X_S(t), X_I(t)) = 1 \ \text{ and}$$

$$\mathbb{P}(R(t) \neq X_I(t) \mid X_S(t), X_I(t) = 0.$$

Hence,

$$\mathbb{P}\big(R(t), I(t) \mid X_S(t), X_I(t)\big) = \mathbb{P}(I(t) \mid X_S(t), X_I(t)).$$

Therefore, the dynamics driving and feasibility constraints can be simplified. This can be done as follows

$$R(t) = X_I(t) \quad \text{and} \quad R(t) = X_I(t) - X_I(t+\Delta t) + X_S(t) - X_S(t+\Delta t)$$

So

$$\begin{aligned} X_I(t) = X_I(t) - X_I(t+\Delta t) + X_S(t) - X_S(t+\Delta t) \Rightarrow \\ -X_I(t+\Delta t) + X_S(t) - X_S(t+\Delta t) = 0 \Rightarrow \\ X_I(t+\Delta t) = X_S(t) - X_S(t+\Delta t). \end{aligned}$$

Hence,

$$I(t) = X_S(t) - X_S(t+\Delta t) = X_I(t+\Delta t), \quad t \in T \tag{5.12}$$

$$0 \leq X_S(t) - X_S(t+\Delta t) \leq X_S(t), \quad t \in T. \tag{5.13}$$

The transition probability of the Markov chain is now obtained by using the constraints, the state space

$$S_{(X_S(t),X_I(t))} = \{(x_S, x_I) \in \mathbb{N}^2 \mid 0 \leq x_S \leq X_S(t), 0 \leq x_I \leq X_S(t), x_S + x_I = X_S(t)\}$$

and the fact that $I(t) = X_I(t+\Delta t)$ (5.12). Therefore, the transition probabilities of the Markov chain are

$$\begin{aligned} &\mathbb{P}(R(t), I(t) \mid X_S(t), X_I(t)) \\ &:= \begin{cases} \mathbb{P}(I(t) = x_I \mid X_S(t), X_I(t)) \quad \text{for} \quad 0 \leq x_S \leq X_S(t), 0 \leq x_I \leq X_S(t), x_S + x_I = X_S(t), \\ 0 \quad \text{otherwise} \end{cases} \end{aligned} \tag{5.14}$$

which we deduced earlier.

Again, we will consider the two possible interventions or actions to control the spread of the infectious disease. Those interventions were vaccination and transmission-reducing interventions. For the purpose of illustration, we will assume that no vaccines are available during the epidemic, so the stationary health policy π only specifies the optimal transmission-reducing intervention $a_t^* \in A_T$, $t \in T$. Further, we also assume that there is only one kind of transmission-reducing intervention available, thus $A_T = \{0,1\}$. The probability that the epidemic will be in state j at decision epoch $t + \Delta t$ given that the epidemic is in state s at decision epoch t and the decision maker chooses action $a \in 0,1$ at decision epoch t is denoted by $p(j \mid s,a)$. The optimal health policy during the epidemic is now obtained by solving the following set of recursive equations or optimality equations

$$v^*(s) = \max_{a \in 0,1} \{r_t(s,a) + \gamma \sum_{j \in S} p(j \mid s,a) v^*(j)\} \quad \text{for} \quad s \in S. \tag{5.15}$$

It can be shown that the solution of the set of equations is given by

$$a^*(s) = \operatorname*{argmax}_{a \in 0,1} \{r_t(s,a) + \gamma \sum_{j \in S} p(j \mid s,a) v^*(j)\}. \tag{5.16}$$

In the following section, we will discuss the application of partially observable Markov decision processes in the medical sciences, specifically for the treatment of ischemic heart disease.

5.2 Ischemic Heart Disease

In this section, we will discuss the application of partially observable Markov decision process for the treatment of ischemic heart disease. First, we will give an explanation of ischemic heart disease. Secondly, we will formulate the partially observable Markov decision process for this specific problem. Ischemic heart disease or coronary artery disease means that the heart is not getting enough oxygen and is often caused by narrowing of the coronary arteries. Ischemic heart disease is a progressive disease and tends to worsen over time. The decision maker or in this case the physician has different options to intervene (actions), namely do nothing, treatment with medication, surgery or perform more test in order to get more information about the status of the disease. These interventions differ in cost, where the cost stands for economic cost, quality of life and invasiveness of procedures. The goal is to develop a strategy that would minimize the expected cost of the treatment. As we have seen in chapter four, a partially observable Markov decision process consists of seven elements: decision times, states, actions, rewards, transition probabilities, observations and observation probabilities. We will now discuss those elements. **Decision times** are given points in time, where decisions are made by the decision maker or in this case a physician. We denote T as the set of decision times and we assume that the set T is discrete. We make this assumption, because we are only interested in discrete-time partially observed Markov decision processes. Further, we denote the elements of the set T as t, which we refer to as time t. Secondly, a **state** is defined as the state of a patient at any point in time. The set of all possible states is denoted by S. An element of the set is denoted by s. The different states for ischemic heart disease are:

- Coronary artery disease: normal, mild-moderate, severe
- Ischemia level: no ischemia, mild-moderate, severe
- Acute myocardial infarction (heart attack): true, false
- Decreased ventricular function: true, false
- Chest pain: no pain, mild, severe
- EKG ischemia: true, false
- Stress test: not available, negative, positive

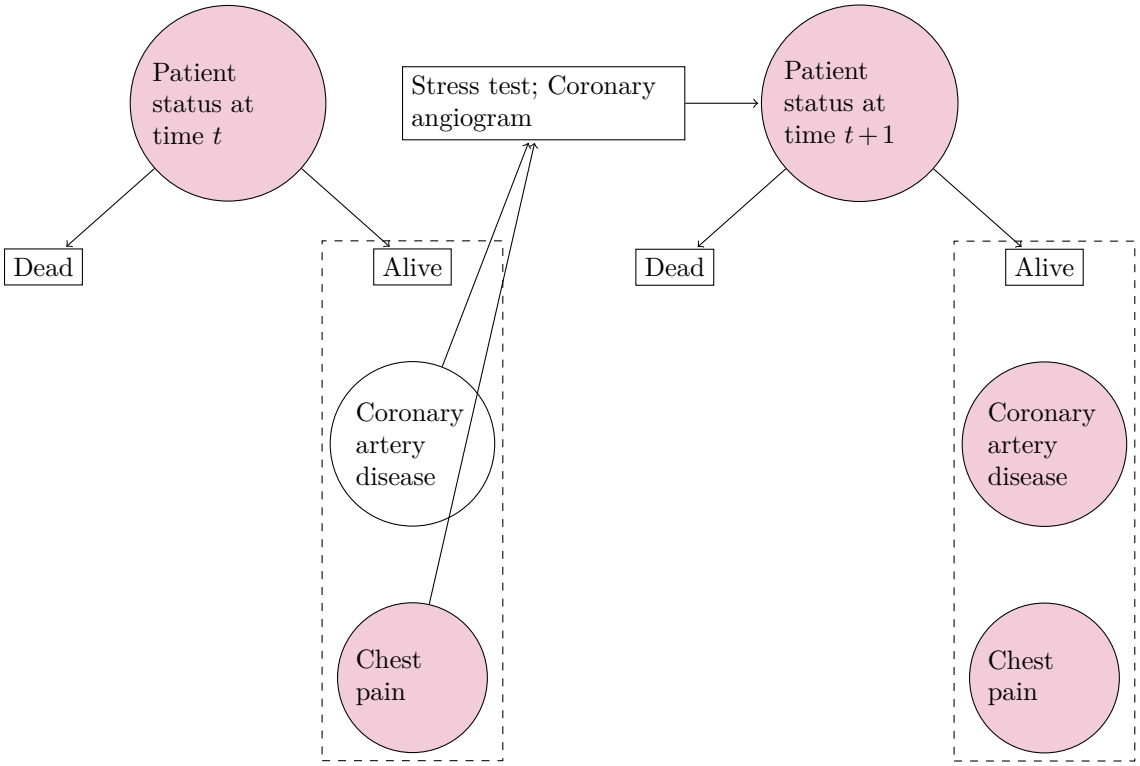
These states are only valid for a patient that is alive. This structure is a bit different then we have seen, since it is hierarchically structured. The state variables are capable to provide more detailed description of the patient state only when the patient is still alive. Further, the states described above represent also the **observations**, since the states are not necessary observable. For example, a patient with severe chest pain has not per definition ischemic heart disease. So severe chest pain is perfectly observable, whether ischemic heart disease is not. We denote the set of all observations by O. The elements of the set of observations will be denoted by o. In the case of ischemic heart disease, an **action** is defined as a treatment or investigative procedure. The set of all possible actions is denoted by A_s. The different actions are:

- No action
- Medication treatment
- Angioplasty; procedure to widen narrowed arteries
- Coronary artery bypass graft surgery
- Stress test (investigative procedure)
- Coronary angiogram (investigative procedure)

Further, the **reward** or cost refer to the economic cost and the physical cost, such as the discomfort of a patient. The reward obtained in state j given state $s \in S$ and action $a \in A_s$ is denoted

by $r_t(j,a,s) = r_t(j) + r_t(a)$, where $r_t(j)$ stands for the costs associated only with the state of a patient and $r_t(a)$ stands for the costs associated with action $a \in A_s$. Lastly, we will discuss transition and observation probabilities. The **transition probability function** $p_t(j \mid s,a)$ denotes the probability that the process will be in state j at the next decision epoch $t+1$, when the decision maker/physician chooses an action $a \in A_s$ in state $s \in S$ at decision time t. In medical terms, the transition probability denotes the state of a patient after a specific treatment. The same is true for the observation probability. An **observation probability** $p_O(o \mid s,a)$ denotes the probability that the decision maker/physician observes a certain observation. Now, we will discuss a simple example.

Example 5.2.1. Suppose a physician sees a patient at time t. The physician observes the patient's status: alive or dead. Assume that the patient is alive and has severe chest pain due to an underlying condition. The physician can observe the chest pain, but the underlying condition, coronary artery disease, is hidden. This is shown in the figure, the purple circles stands for observable states and the white circles stands for hidden states. When the patient is alive there are multiple states and observations possible, this is shown inside the dashed box. After observing the chest pain, the physician decides to employ investigative procedures or actions. These actions are a stress test and a coronary angiogram. After doing a stress test and a coronary angiogram, the process evolves to decision time $t+1$. The physician sees the patient again at decision time $t+1$. He now observes chest pain and coronary artery disease due to the investigative procedures he employed at time t.



♦

We have now formulated the partially observable Markov decision process for the treatment of ischemic heart disease and discussed an example. Once a partially observable Markov decision process is defined it could be converted into a **belief state Markov decision process**. We have not discussed belief state Markov decision processes before, since it is a bit beyond the scope of this thesis. However, we will give a short motivation on belief state Markov decision process such that we can apply it later in our model for the treatment of ischemic heart disease.

A **belief state** b assigns a probability to all states $s \in S$. The probability that the process is in state s is denoted by $b(s)$, with $\sum_{s \in S} b(s) = 1$. The probability of observing a certain observation $o \in O$ given action $a \in A_s$ and belief state b is given by

$$p(o \mid b,a) = \sum_{j \in S} p_o(o \mid j,a) \sum_{i \in S} p_t(j \mid i,a) b(i), \tag{5.17}$$

where $i \in S$ is the current state, $j \in S$ the next state and $b(i)$ the probability that the process is in state i. We can now update our belief, this means that we can formulate the next state belief, which is acquired after observing a certain observation $o \in O$ and choosing action $a \in A_s$. Therefore, the **updated belief** b' is defined as

$$b'(i) = \frac{p_o(o \mid i,a) \sum_{j \in S} p_t(i \mid j,a) b(j)}{p(o \mid b,a)} \quad \text{and} \quad p(o \mid b,a) > 0, \tag{5.18}$$

where $i \in S$ is the current state, $j \in S$ the next state and $b(j)$ the probability that the process is in state j. Without any proof, we will state the Bellman equations for a belief state Markov decision process. The optimality equations or Bellman equations are given by

$$\upsilon^*(b) = \max_{a \in A_s} \{ r_t(b,a) + \gamma \sum_{o \in O} p(o \mid b,a) \upsilon^*(b') \},$$

where b is a belief state, b' updated belief state, $r_t(b,a)$ the expected reward and $\gamma \in [0,1]$ the discount factor.

If we apply this theory to our model for the treatment of ischemic heart disease, then this will lead to two improvements. The first improvement is that not all information is needed, so we can work with less information and still manage to solve the problem. This comes from the fact that not all state variables at a given time are necessary to define the belief state. It is enough to use a belief state that is defined only over the state variables that are directly used in the transition from one decision epoch to the other. So in example (5.2.1), it suffices to use only the state variables patient status, coronary artery disease and chest pain in stead of all possible states in order to define the belief state. Those variables are called **information state variables** and are denoted by d. The second improvement is using a so-called hybrid information state. A hybrid information state $\{o_d, b_d\}$ consists of two components, namely a vector of observable information states o_d and a vector of belief information states b_d. This is a improvement, since the decision maker can work with the actual value of a state when it is perfectly observed rather than the beliefs. So in case that some variables are perfectly observable and others are hidden, we would like to work with hybrid information states. Hence, a belief state Markov decision process has some improvements over a partially observable Markov decision process. The next chapter will contain the conclusion of this thesis.

Conclusion

In this thesis, we explained Markov decision processes and partially observed Markov decision processes. We started with the explanation of a sequential decision process and after that we elaborated on Markov chains. In our elaboration on Markov chains, we discussed the Markov property, the Chapman-Kolmogorov equations and various properties associated with the classification of states. Further, in our explanation of Markov decision processes, we discussed the five components, decision times, states, actions, transition probabilities and rewards, decision rules and policies. Also, we explained the finite-horizon and the infinite-horizon Markov decision processes. In our elaboration on partially observed Markov decision processes, we discussed the seven elements, which were decision times, states, actions, observations, transition probabilities, observation probabilities and rewards. We have seen that Markov decision processes and partially observed Markov decision processes differ in observability. A Markov decision process is perfect observed, so the decision maker has all information available at all times. However, a partially observed Markov decision process is not perfect observed. In this case, the decision maker has not all the information, so he has to make his decision based on observations and previous actions. Lastly, we showed some applications of Markov decision processes and partially observed Markov decision processes to medical sciences, in particular the spread of infectious disease and the treatment of ischemic heart disease. We needed to make some simplifying assumptions in order to apply Markov decision processes in medical sciences.

Bibliography

[1] A.R. Cassandra *Optimal Policies for Partially Observable Markov Decision Processes*
February 27, 1995

[2] F.M. Dekking et al. *A Modern Introduction to Probability and Statistics: understanding why and how*
Springer; 1st edition reprint (October, 2010), ISBN: 978-1-84996-952-9

[3] M. Hauskrecht, H. Fraser *Planning treatment of ischemic heart disease with partially observable Markov decision processes*
Artificial Intelligence in Medicine 18 (2000) 221–244

[4] V. Krishnamurthy *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*
Cambridge University Press; 1st edition (2016), ISBN 978-1-107-13460-7

[5] V.G. Kulkarni *Modeling and Analysis of Stochastic Systems* CRC Press; 2nd edition (2009), ISBN 978-1-4398-0877-1

[6] M.L. Puterman *Markov Decision Processes: Discrete Stochastic Dynamic Programming*
John Wiley & Sons, inc.; 2nd edition (February, 2005), ISBN-13: 978-0471727828

[7] S.M. Ross *Introduction to Probability Models*
Academic Press; 10th edition (December, 2009), ISBN: 978-0-12-375686-2

[8] A. Schaefer et al. *Modeling medical treatment using markov decision processes*
Vol. 23. Springer; New York: 2005. Operations Research and Health Care - A Handbook of Methods and Applications; p. 593-612.

[9] J. Taylor *Markov Decision Processes: Lecture Notes for STP 425*
November 26, 2012

[10] R. Yaesoubi, T. Cohen *Dynamic Health Policies for Controlling the Spread of Emerging Infections: Influenza as an Example*
(2010)

[11] R. Yaesoubi, T. Cohen *Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies*
European Journal of Operational Research. September 6, 2011