# Markov Decision Processes with Information Costs

Theory and Application

Dissertation
Fachbereich Mathematik und Informatik
Freie Universität Berlin

**Stefanie Winkelmann**

April 2013

Betreuer: Prof. Dr. Christof Schütte
Freie Universität Berlin
Fachbereich Mathematik und Informatik
Arnimallee 2-6
14195 Berlin

Gutachter: Prof. Dr. Christof Schütte
Prof. Dr. Michael Dellnitz (Universität Paderborn)

Tag der Disputation: 5. Juli 2013

# Contents

# Introduction

The theory of Markov decision processes (MDP) is a well established tool for analyzing situations in which the dynamics of a stochastic process can be influenced by a decision maker. It provides a framework for solving optimization problems that arise in a wide range of fields like operations research, epidemic control or management science [74]. The research of Markov decision processes goes back to the 1950s. Of central relevance was the introduction of the *dynamic programming* concept by Bellman [5,6]. It allows to break down a complex problem into smaller subproblems and is expressed by the so called *Bellman equation* which states the optimization problem in a recursive form.

A Markov decision process is a Markov process which is subject to the control of a decision maker. Given the *state* of the process at some point in time, the decision maker has to choose a suitable *action* which regulates the future stochastic dynamics of the process. Depending on the process evolution, the control can be adapted anytime. The controlled process produces costs according to a given cost function, and the goal is to find a control *policy* which minimizes a given *cost criterion*. The precise description and solution of a Markov control problem depend on the characteristics of the underlying state space and the time index which both can be discrete or continuous. Another basic component of a Markov control model is the observability of the process: in the original Markov control theory the process is assumed to be observable at all times, while in the theory of *partially observable Markov decision processes* and related models the information about the process is incomplete. In this thesis we will focus on **continuous-time Markov decision processes on discrete state spaces** and develop a new model for the situation of **limited state information**.

In the general theory of Markov decision processes, the time scale of the control procedure coincides with the one of the process itself, i.e. a continuous-time process allows for a continuous interaction by the controller. This is based on the assumption that the information about the process (even if it is incomplete) is provided permanently over time, and that, given a new information, the action can immediately be adapted. However, in many applications (like medical therapies, asset management, machine maintenance problems etc., see e.g. [35, 36, 51]) such a permanent information and control is not feasible. Instead, it is an obvious idea that the state of a continuous-time process might only be determinable at single points in time - just as the health status of a patient has to be determined by a medical test, or the condition of a machine is identified by an inspection. Moreover, such a test or inspection is in general not free of charge, but produces a particular instantaneous cost, which is independent of the process costs measured by the given cost function. The Markov control model presented in this thesis will contain an explicit parameter

$$k_{\text{info}} > 0$$

of *information costs* which arise each time the state of the process is determined and the action is adapted. While the controlled process itself is continuous in time, its observation and the choice of actions take place at **discrete but flexible points in time**. To determine these *observation times* will be part of the control procedure. That is, in the new setting a control policy will assign to each state not only an action but also a date for the next state observation. We will reformulate the common criteria of *discounted costs* and *average costs* such that they include the cost of information which arise during the control procedure, and derive the corresponding dynamic programming equations which will be the basis for an efficient numerical calculation of the related optimal control policies.
A detailed analysis of the new Markov control problem with information costs will be followed by an application in the medical context where we determine optimal therapeutic strategies for the treatment of the human immunodeficiency virus (HIV) in resource-rich and resource-poor settings.

The problem of a Markov decision process which is not completely observable is well-known in the literature. The most general approach to handle such problems of limited information is given by the theory of partially observable Markov decision processes (POMDP), see [10, 33, 46, 49, 63] among others. Here, the process is usually assumed to be discrete in time, and the degree of information is determined by the chosen action. Based on indirect observations, the process is turned into a completely observable process on the set of probability distributions. On the one hand, this allows a very general analysis and application within the setting of discrete time; on the other hand, due to the high dimensionality of the new state space of distributions, the resulting optimization problem is very complex to solve.
More special settings of Markov control with incomplete state information, both for discrete and for continuous time, are given in the context of machine maintenance. These approaches all have in common that they assume the controlled dynamics to exhibit a special structure: interaction consists of sending the process to a special state (e.g. repairing a machine); the state space exhibits an ordered structure; the dynamics are in some sense monotone; or other assumptions [1, 34, 50, 51, 54, 72].
The model presented in this thesis will be suitable for a general discrete state space and arbitrary continuous dynamics, without fixing any interpretation or systematic. In contrast to the theory of POMDPs, the information about the state of the process will be independent of the applied action. We will assume that the state determination always delivers perfect information about the state of the process at the observation times and that the action can only be adapted after such a state obser-

vation (and not “blindly“ in between). This permits a compact and comprehensible formulation of the control problem, as well as an efficient calculation of its solution by standard numerical methods. At the same time, it delivers a straightforward interpretation which is suitable for many real-world applications.

## Outline of the thesis

In Chapter 1 we will give an introduction to the theory of Markov decision processes and its modifications, presenting the standard of knowledge on the basis of the existing Markov control literature. We begin in Section 1.1 with completely observable Markov decision processes in continuous time and analyze the two cost criteria of discounted costs and average costs over an infinite time horizon. For both criteria, the main result is given by the Bellman equation which does not only characterize the function of optimal costs, but also delivers the basis for a numerical calculation of the optimal policy. A detailed description of the existing optimization algorithms will be given in Section 1.2. Then, in Section 1.3, we will see how the situation of incomplete state information is handled in the literature, describing both the general theory of partially observable Markov decision processes, as well as a range of special settings. This first chapter is meant to provide a scientific background for the analysis that will follow in Chapter 2. Moreover, some of the presented results can directly be transferred to the new setting, which will simplify the further analysis.

Given this overview, we will develop the novel Markov control model containing the additional parameter $k_{\text{info}} > 0$ of information costs (Chapter 2). In Section 2.1 we will explain the corresponding control procedure in detail and define the two criteria of discounted costs and average costs within the new setting. An extensive analysis of the two criteria will be given in the subsequent Sections 2.2 (discounted costs) and 2.3 (average costs). In each case, the first step is to reformulate and prove the Bellman equation, which will be done in Section 2.2.1 resp. 2.3.1. Interestingly, the approaches will be completely different: While the discounted-cost criterion permits a direct analysis, the average-cost criterion can be handled by turning the given process into an equivalent completely observable Markov decision process which allows to transfer the results presented in Section 1.1 to our new framework. For both criteria, we will calculate a *cost splitting* in order to determine the contribution of the information costs to the total costs (Section 2.2.2 resp. 2.3.2). At the same time, this splitting allows to make an unbiased comparison with the optimal costs of the related original Markov control problem. We will further analyze the impact of the information cost parameter $k_{\text{info}}$ on the optimal costs and on the optimal control policy, exploring properties of monotony and continuity to some extend (Section 2.2.3 resp. 2.3.3). The underlying idea is that a reduction of the information costs $k_{\text{info}}$ should lead to more frequent testing and a decrease in the process costs. Again, we can link the new model to the original Markov control model by considering the limit $k_{\text{info}} \to 0$. Another question that will be answered for both criteria concerns the optimal observation times given a fixed $k_{\text{info}}$ (Section 2.2.4 resp. 2.3.4): How do deviations from the optimal observation times affect the costs? This question has a practical motivation, thinking of circumstances that prevent from an exact adherence of the prescribed test dates in real-world applications. We close this chapter by comparing the two cost criteria with each other (Section 2.4).

In Chapter 3 we will apply the developed theory of Markov control with information costs to an example in a medical context. Considering the process of drug resistance development of the human immunodeficiency virus (HIV), we determine optimal therapeutic strategies for resource-rich and resource-poor settings from a national economic perspective. In this application, the cost of information $k_{\text{info}}$ refers to the price of a drug resistance test.
The first step will be to define all parameters of the corresponding Markov control model (Section 3.1). Given the model, we will determine the optimal control policy for both settings and calculate the cost splitting for different values of information costs $k_{\text{info}}$ (Section 3.2). We will use the cost splitting for a comparison with the two extreme situations of vanishing information costs (as in the original Markov control theory) and infinite information costs (resulting in constant control without state observations). In Section 3.3 we will analyze the sensitivity with respect to deviations from the optimal testing dates for the resource-poor setting where it may not be possible to follow a recommended diagnostic surveillance scheme accurately due to a limited infrastructure. Finally, Section 3.4 will deal with the life expectancy of a patient under optimal therapy, investigating which impact the reduction of information costs and treatment costs have in this matter.

## Acknowledgements

I would like to thank all those people who supported and encouraged me during the last years of research. First of all, I thank my advisor Christof Schütte for giving me the opportunity to write this thesis, guiding my research and giving helpful advice. I am happy to be a member of the biocomputing group which offers such a pleasant and motivating working atmosphere. Especially, I thank Carsten Hartmann for steady exchange and valuable suggestions, Marco Sarich for being a pleasant office mate and giving helpful support in scientific and technical issues, Max von Kleist for a great teamwork, and Natasa Djurdjevac, Antonia Mey and Ralf Banisch for reading this thesis and giving useful comments.

Finally, I would like to thank my family for their steady mental and practical support. The helping hand of my parents and my mother-in-law allowed me to combine the pleasure of motherhood with an interesting and delighting employment. I thank Lars for walking all the way together and giving me strength, and Clara for just being there.

Chapter 1

# Fundamentals of Markov Control Theory

Markov decision processes (MDPs) provide a mathematical framework for modeling situations in which the evolution of a process is partly random and partly controllable. These situations arise in many application areas such as machine maintenance, population control, financial engineering, manufacturing, queuing systems or epidemic control. Within the last century, Markov decision processes have been studied by many authors, see e.g. [5, 24, 26, 32, 43, 48]. One of the pioneers in this area of research was the American mathematician Richard E. Bellman (1920-1984). His central contribution is given by the *Bellman equation* which rephrases an optimization problem in a recursive form and thereby breaks it down into smaller subproblems [5, 6]. This is the basic step for an efficient determination of the optimal solution by numerical methods.

In this chapter we will present the main ideas and results for Markov decision processes on a countable state space based on standard Markov control literature. Our aim is not to deliver a complete and technically general analysis, but to give an overview which allows to put the new ideas of Chapter 2 into the right context. In this regard, the observability of the process will be of main interest, and so we will focus not only on the original Markov control theory (where the process is assumed to be observable at all times), but also give an introduction to the theory of partially observable Markov decision processes (POMDPs) and related approaches.

In Section 1.1 we will consider completely observable Markov decision processes on a discrete state space which are continuous in time. We will present the two common criteria of *discounted costs* and *average costs* for an infinite time horizon and formulate the related optimality equations. Sometimes we will restrict the analysis to finite state spaces in order to permit compact formulations and justifications which are not taken from the literature but are based on our own approaches. For the numerical implementation which will be the object in Section 1.2 such a limitation to finite state spaces is naturally given, anyway. Here, we will present different algorithms for the calculation of the optimal control policy for both cost criteria. In Section 1.3, we will briefly introduce the theory of POMDPs and see how the situation of limited information is handled in the literature.

## 1.1 Original Markov Control Theory

In this section, we consider continuous-time Markov decision processes which are at all times completely observable. The state space is assumed to be discrete, such that, for a fixed control rule, the resulting stochastic process is a Markov jump process. Following common practice, the first step is to identify a *Markov control model* containing all relevant parameters of the control process. Based on such a Markov control model, the term *policy* – as a rule for making the decisions – can accurately be defined in a mathematical sense. We will see how the controlled process evolves over time and formulate the two optimality criteria of discounted costs and average costs. Both criteria will be analyzed subsequently, discovering the structure of the cost functional for a given policy and formulating the corresponding Bellman equation in terms of the *value function* of optimal costs.
In this section, we closely follow the approaches presented in [26] – a book by X. Guo and O. Hernández-Lerma (2009) which is specifically devoted to continuous-time Markov decision processes.

### The Markov control model

A Markov control model is given by a tuple

$$\Big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{L_a : a \in \mathcal{A}\},\, c\Big) \tag{1.1}$$

with the following components [26].

- The *state space* $\mathcal{S}$: This is the set of all possible states of the considered process. We assume $\mathcal{S}$ to be denumerable.
- The *action space* $\mathcal{A}$: This is the set of actions which are available in order to control the process. We allow $\mathcal{A}$ to be any topological space and denote the corresponding Borel $\sigma$-algebra by $\mathcal{B}(\mathcal{A})$.
- The family $\{\mathcal{A}(x) : x \in \mathcal{S}\}$ of *available actions*: For each $x \in \mathcal{S}$, it is $\mathcal{A}(x) \subset \mathcal{A}$ the set of actions that are available to the controller when the process is in state $x$.
- The set of *generators* $(L_a)_{a\in\mathcal{A}}$: For each action $a \in \mathcal{A}$, the infinitesimal generator $L_a$ describes the dynamics of the process given this action. More precisely, $L_a(x,y) \geq 0$ is the transition rate for a transition from $x \in \mathcal{S}$ to $y \in \mathcal{S}$, $y \neq x$, while $L_a(x,x)$ is defined by

$$L_a(x,x) := -\sum_{y\neq x} L_a(x,y).$$

We assume the transition rates to be stable in the sense of

$$\sup_{a\in\mathcal{A}(x)} l_a(x) < \infty \quad \forall x \in \mathcal{S},$$

where $l_a(x) := -L_a(x,x)$.

- The *cost function* $c : \mathcal{S} \times \mathcal{A} \to [0, \infty)$: Depending on the actual state and the chosen action this function measures the costs produced by the process per unit of time.

We will mostly be concerned with the situation where $\mathcal{A}(x) = \mathcal{A}$ for all $x \in \mathcal{S}$, i.e. for every state all actions are available.
Given the Markov control model (1.1), the control procedure consists of assigning to each state $x \in \mathcal{S}$ an action $a \in \mathcal{A}$ which is applied whenever the process is in this state. This motivates the following definition.

**Definition 1.1** (Markov policy)**.** A *randomized, time-dependent Markov policy* is a set of functions $(\pi_t)_{t\geq 0}$,

$$\pi_t : \mathcal{S} \times \mathcal{B}(\mathcal{A}) \to [0, 1],$$

which satisfies the following conditions.

(i) For all $x \in \mathcal{S}$ and $B \in \mathcal{B}(\mathcal{A})$, the mapping $t \mapsto \pi_t(x, B)$ is measurable on $[0, \infty)$.

(ii) For all $t \geq 0$ and $x \in \mathcal{S}$, the mapping $B \mapsto \pi_t(x, B)$ is a probability measure on $\mathcal{B}(\mathcal{A})$ with $\pi_t(x, \mathcal{A}(x)) = 1$. It is $\pi_t(x, B)$ the probability that an action $a \in B$ is chosen when the process is in state $x \in \mathcal{S}$ at time $t \geq 0$.

Such a policy is called *stationary* if $\pi_t(x, B) \equiv \pi(x, B)$ is independent of $t$. It is called *deterministic* if for each $t \geq 0$ and $x \in \mathcal{S}$ there is an $a \in \mathcal{A}(x)$ such that $\pi_t(x, \cdot) = \delta_a$ is a Dirac measure. A deterministic stationary Markov policy is given by a function

$$u : \mathcal{S} \to \mathcal{A}$$

with $u(x) \in \mathcal{A}(x)$ for all $x \in \mathcal{S}$, assigning to each state an available action in a deterministic way.
We denote the set of all randomized, time-dependent Markov policies by $\Pi$ and the set of all deterministic stationary Markov policies by $\mathcal{U}$.

Here, the term “Markov” refers to the fact that the policy is a function of the actual state $x \in \mathcal{S}$ and does not depend on the complete history of the process.

### The continuous-time Markov decision process

A policy $\pi \in \Pi$ determines the evolution of the control system in the following way. Given that the process is in state $x \in \mathcal{S}$ at time $t \geq 0$, the decision maker chooses an action $a \in \mathcal{A}$ according to the distribution $\pi_t(x, \cdot)$. During the time interval $[t, t + dt)$ the process then evolves according to the generator $L_a$ and produces costs $c(x, a)$ per unit of time. For a deterministic stationary policy $u \in \mathcal{U}$ the action remains constant until a transition to a state $y \in \mathcal{S}$ with $u(y) \neq u(x)$ occurs, see Figure 1.1.

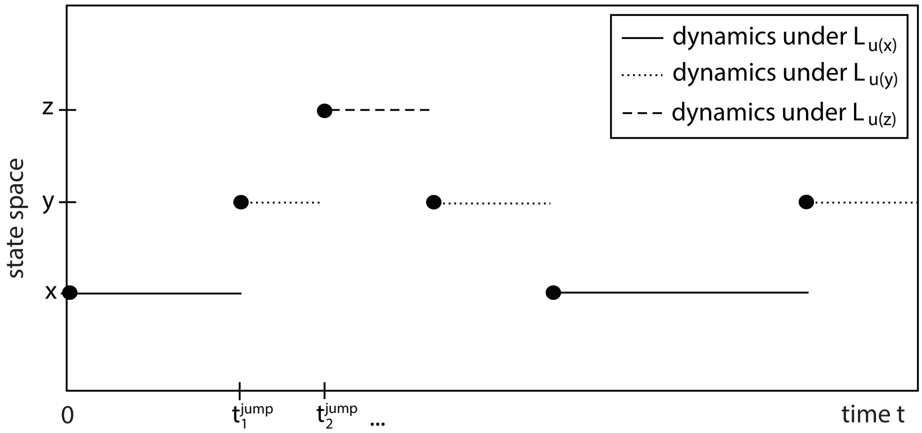

Figure 1.1: **Controlled Markov process.** Possible trajectory of a Markov decision process given a deterministic stationary policy $u \in \mathcal{U}$. Starting in a state $x \in \mathcal{S}$, the dynamics of the process are given by the generator $L_{u(x)}$, i.e. the process stays in $x$ for some random period of time which is exponentially distributed with parameter $l_u(x)$ and then jumps (at time $t_1^{\text{jump}}$) to a state $y \in \mathcal{S}$ with probability $\frac{L_{u(x)}(x,y)}{l_u(x)}$. As soon as the jump occurs, the action is adapted and the process proceeds according to the generator $L_{u(y)}$.

More precisely, a policy $\pi \in \Pi$ together with an initial distribution $\nu$ on $\mathcal{S}$ defines a probability measure $\mathbb{P}_\nu^\pi$ on the set of possible state-action-realizations

$$\{(X_t, A_t)_{t\geq 0} : \; X_t \in \mathcal{S}, A_t \in \mathcal{A} \; \forall t \geq 0\}$$

by

- $\mathbb{P}_\nu^\pi(X_0 = x) = \nu(x)$,
- $\mathbb{P}_\nu^\pi(A_t \in B | X_t = x) = \pi_t(x, B)$,
- $\frac{\partial}{\partial t}\mathbb{P}_\nu^\pi(X_t = x | A_t = a) = \sum_{y \in \mathcal{S}} L_a(y,x)\mathbb{P}_\nu^\pi(X_t = x | A_t = a)$,

where $t \geq 0$, $B \in \mathcal{B}(\mathcal{A})$, $x \in \mathcal{S}$ and $a \in \mathcal{A}$.
For the case of $\nu = \delta_x$ (deterministic start in $x \in \mathcal{S}$) we write $\mathbb{P}_x^\pi$. The corresponding expectation value is denoted by $\mathbb{E}_\nu^\pi$ resp. $\mathbb{E}_x^\pi$.

In this thesis, we will restrict the analysis to deterministic stationary policies $u \in \mathcal{U}$. This does not mean any crucial restraint: Given any measurable function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we can set

$$f(x, \pi) := \int_{\mathcal{A}(x)} f(x,a)\pi(x, da)$$

in order to extend the definition to non deterministic policies $\pi \in \Pi$, which allows a direct transfer of all analytic results to the more general set of policies $\Pi$. Moreover,

for the cost criteria that will be considered (see page 10) the optimal policy – assuming its existence – is deterministic and stationary anyway [26]. The advantage of the restriction is that analytic expressions get more compact and, at the same time, allow for a straightforward interpretation.
Given a policy $u \in \mathcal{U}$, we write $\mathbb{E}^u_\nu$ resp. $\mathbb{E}^u_x$ for the expectation value referring to the measure $\mathbb{P}^u_\nu$ resp. $\mathbb{P}^u_x$ which has the same definition as $\mathbb{P}^\pi_\nu$ resp. $\mathbb{P}^\pi_x$. Especially, it holds

$$\mathbb{P}^u_\nu(A_t \in B | X_t = x) = \delta_{u(x)}(B).$$

In order to illustrate the given situation, we consider the following simple example which will reappear several times throughout the entire work.

**Example 1.2** (Two states)**.** *Let $\mathcal{S} = \{x_1, x_2\}$ and $\mathcal{A} = \mathcal{A}(x_1) = \mathcal{A}(x_2) = \{a_1, a_2\}$ with*

$$L_1 = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix}, \quad L_2 = \begin{pmatrix} -0.1 & 0.1 \\ 0.1 & -0.1 \end{pmatrix}$$

*and $c(x,a) = c_\mathcal{S}(x) + c_\mathcal{A}(a)$, where $c_\mathcal{S}(x_1) = 0$, $c_\mathcal{S}(x_2) > 0$, $c_\mathcal{A}(a_1) = 0$, $c_\mathcal{A}(a_2) > 0$. In this setting, $x_1$ is the "good" state: As long as the process is in state $x_1$, it does not produce any costs, while when being in state $x_2$, it produces costs at rate $c_\mathcal{S}(x_2) > 0$. The first action is free of charge ($c_\mathcal{A}(a_1) = 0$), but has a small rate to push the process back to state $x_1$ when being in state $x_2$, while the costly action $a_2$ increases this rate.*
*A possible policy consists of choosing the free action $a_1$ as long as the process is in the cost-free state $x_1$, while choosing action $a_2$ when the process is in state $x_2$ in order to quickly push it back to state $x_1$. Whether this policy is in fact favorable, can be answered after formulating an optimization criterion.*

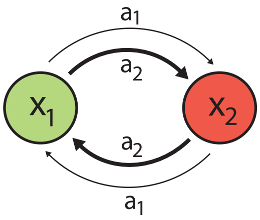

Figure 1.2: **2-state-example.** Action $a_2$ increases the transition rates between the "good" state $x_1$ and the "bad" state $x_2$, which is indicated by the thickness of the corresponding arrows.

### Optimality criteria

The aim of the decision maker is to find a policy which optimizes a given performance criterion. For an infinite time horizon there are mainly two common criteria, namely

a) The *expected discounted-cost criterion*: Given an initial state $x \in \mathcal{S}$ and a discount factor $\lambda > 0$, the total expected discounted costs under control $u \in \mathcal{U}$ are defined by

$$J_\lambda(x, u) := \mathbb{E}_x^u \left( \int_0^\infty e^{-\lambda s} c(X_s, u(X_s))\, ds \right). \tag{1.2}$$

The corresponding *optimal discounted-cost function* or *value function of discounted costs* is given by

$$V_\lambda(x) := \inf_{u \in \mathcal{U}} J_\lambda(x, u).$$

b) The *expected average-cost criterion*: Given an initial state $x \in \mathcal{S}$, the long-run expected average costs under control $u \in \mathcal{U}$ are defined by

$$\bar{J}(x, u) := \limsup_{T \to \infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T c(X_s, u(X_s))\, ds \right). \tag{1.3}$$

The corresponding *optimal average-cost function* or *value function of average costs* is given by

$$\bar{V}(x) := \inf_{u \in \mathcal{U}} \bar{J}(x, u).$$

Given $\varepsilon > 0$, a policy $u^* \in \mathcal{U}$ is said to be $\varepsilon$*-optimal* (with respect to discounted resp. average costs) if

$$J_\lambda(x, u^*) \leq V_\lambda(x) + \varepsilon \quad \forall x \in \mathcal{S}$$

resp.

$$\bar{J}(x, u^*) \leq \bar{V}(x) + \varepsilon \quad \forall x \in \mathcal{S}.$$

The policy $u^*$ is called *optimal* (with respect to discounted resp. average costs) if

$$J_\lambda(x, u^*) = V_\lambda(x) \quad \forall x \in \mathcal{S}$$

resp.

$$\bar{J}(x, u^*) = \bar{V}(x) \quad \forall x \in \mathcal{S}.$$

In order to guarantee the existence of an optimal policy, one has to make further assumptions (for example that $\mathcal{A}$ is finite or that $\mathcal{A}$ is compact and $c$ and $L$ are continuous on $\mathcal{A}$). While the existence of an optimal policy cannot be taken for granted, an $\varepsilon$-optimal policy exists without any further assumptions.

The set of policies is a very high-dimensional space such that finding an optimal policy is a nontrivial problem. However, as we will see, the optimal policy can be characterized by a dynamic programming equation, the so called Bellman equation, which allows an efficient numerical computation.

**Remark 1.3** (Time-discrete setting)**.** *The introduced Markov control model* (1.1) *refers to a continuous-time setting which will be relevant in Chapter 2. However, for some of the considerations in the following sections we will have to deal with a discrete time parameter. For this reason, we here briefly present the discrete-time analogue of the Markov control model, compare [41].*
*A discrete-time Markov control model is given by a tuple*

$$\big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{P_a : a \in \mathcal{A}\},\, \tilde{c}\,\big), \tag{1.4}$$

*where $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{A}(x)$ are defined as before. For each $a \in \mathcal{A}$, $P_a(x,y) = \mathbb{P}(X_{n+1} = y|X_n = x, A_n = a)$ is the probability for a transition from $x \in \mathcal{S}$ to $y \in \mathcal{S}$ given that action $a$ is applied, and $\tilde{c} : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ is a cost function giving the costs per (discrete) time step.*
*Just as in the continuous-time situation, a (deterministic stationary) policy is a function $u : \mathcal{S} \to \mathcal{A}$ with $u(x) \in \mathcal{A}(x)$ for all $x \in \mathcal{S}$ indicating for each state which action to choose, and such a policy (together with an initial distribution) defines a probability measure on the set of discrete state-action-combinations*

$$\{(X_n, A_n)_{n\in\mathbb{N}_0} :\ X_n \in \mathcal{S}, A_n \in \mathcal{A}\ \forall n \in \mathbb{N}_0\}\,.$$

*The expected discounted-cost criterion is given by*

$$\tilde{J}_\lambda(x,u) := \mathbb{E}_x^u\left(\sum_{n=0}^{\infty} e^{-\lambda n}\tilde{c}(X_n, u(X_n))\right),$$

*while the expected average-cost criterion gets the form*

$$\tilde{J}(x,u) := \limsup_{N\to\infty}\ \mathbb{E}_x^u\left(\frac{1}{N}\sum_{n=0}^{N-1}\tilde{c}(X_n, u(X_n))\right).$$

*In the case of ergodic dynamics (compare page 20 for details), the value $\tilde{J}(x,u)$ does not depend on $x \in \mathcal{S}$ such that we can define the constant*

$$\tilde{\eta}_u := \tilde{J}(x,u)$$

*of average costs independent of $x$.*

In the next sections the two cost criteria (discounted costs and average costs) will be analyzed for the continuous-time setting. We will give an overview of the fundamental results based on standard Markov control literature. For both criteria we will proceed in a parallel manner: In a first step, the cost functional $J_\lambda(x,u)$ resp. $\bar{J}(x,u)$ of a given deterministic policy $u$ is analyzed, introducing an appropriate recursion which, as for finite state spaces, uniquely characterizes the functional. In a second step, we consider the value function of optimal costs and present criteria for a policy to be optimal.
By limiting some of the analysis to finite state spaces, we will be able to give our own simplified approach to prove the presented statements. For more general results, we will refer to common Markov control literature, e.g. [26, 48, 75].

### 1.1.1 The Discounted-Cost Criterion

Given the Markov control model (1.1), we consider, for a given policy $u \in \mathcal{U}$, the criterion of expected discounted costs

$$J_\lambda(x,u) = \mathbb{E}_x^u \left( \int_0^\infty e^{-\lambda s} c(X_s, u(X_s))\, ds \right)$$

defined in (1.2). Before turning to the analysis of optimal policies and related minimal costs, we provide an insight into the characteristics of the cost functional $J_\lambda(x,u)$ for an arbitrary policy $u \in \mathcal{U}$.

**Cost functional $J_\lambda(x,u)$ for a given policy $u \in \mathcal{U}$**

For a given policy $u \in \mathcal{U}$, we set

$$l_u(x) := -L_{u(x)}(x,x).$$

If the process is in state $x \in \mathcal{S}$ and policy $u$ is applied, the jump time (i.e. the time until the process leaves $x$ for the first time) is exponentially distributed with parameter $l_u(x)$. By means of this jump time parameter $l_u(x)$ one can formulate a recursive equation for the functional of discounted costs [26].

**Lemma 1.4.** *Given a policy $u \in \mathcal{U}$, the corresponding cost functional $J_\lambda(x,u)$ fulfills the recursion*

$$J_\lambda(x,u) = \frac{c(x,u(x))}{\lambda + l_u(x)} + \frac{1}{\lambda + l_u(x)} \sum_{y \neq x} L_{u(x)}(x,y) J_\lambda(y,u). \tag{1.5}$$

*Proof.* Starting in $x \in \mathcal{S}$, let $t_1$ be the (random) time of the first jump to another state $y \in \mathcal{S}$, $y \neq x$. We split the cost functional $J_\lambda(x,u)$ up into the costs arising before $t_1$ and after $t_1$:

$$\begin{aligned}
& J_\lambda(x,u) \\
= \; & \mathbb{E}_x^u \left( \int_0^{t_1} e^{-\lambda s} c(x,u(x))\, ds + \int_{t_1}^\infty e^{-\lambda s} c(X_s, u(X_s))\, ds \right) \\
= \; & \mathbb{E}_x^u \left( \int_0^{t_1} e^{-\lambda s} c(x,u(x))\, ds \right) + \mathbb{E}_x^u \left( \int_{t_1}^\infty e^{-\lambda s} c(X_s, u(X_s))\, ds \right) \\
= \; & \mathbb{E}_x^u \left( \frac{1 - e^{-\lambda t_1}}{\lambda} \cdot c(x,u(x)) \right) \\
& + \sum_{y \neq x} \mathbb{P}_x^u(X_{t_1} = y) \cdot \mathbb{E}_x^u \left( \int_{t_1}^\infty e^{-\lambda s} c(X_s, u(X_s))\, ds \,\middle|\, X_{t_1} = y \right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} \int_0^\infty l_u(x)e^{-l_u(x)t} \cdot \Bigg[\frac{1-e^{-\lambda t}}{\lambda} \cdot c(x,u(x)) \\
&\qquad + \sum_{y\neq x} \mathbb{P}_x^u(X_{t_1}=y) \cdot \mathbb{E}_x^u \left( \int_{t_1}^\infty e^{-\lambda s} c(X_s,u(X_s))\, ds \,\middle|\, t_1 = t, X_{t_1} = y \right) \Bigg] dt \\
&\overset{(b)}{=} \int_0^\infty l_u(x)e^{-l_u(x)t} \cdot \Bigg[\frac{1-e^{-\lambda t}}{\lambda} \cdot c(x,u(x)) \\
&\qquad + \sum_{y\neq x} \frac{L_{u(x)}(x,y)}{l_u(x)} \cdot e^{-\lambda t} \cdot \mathbb{E}_y^u \left( \int_0^\infty e^{-\lambda s} c(X_s,u(X_s))\, ds \right) \Bigg] dt \\
&= \int_0^\infty l_u(x)e^{-l_u(x)t} \cdot \left[\frac{1-e^{-\lambda t}}{\lambda} \cdot c(x,u(x)) + \sum_{y\neq x} \frac{L_{u(x)}(x,y)}{l_u(x)} \cdot e^{-\lambda t} \cdot J_\lambda(y,u)\right] dt \\
&= \frac{c(x,u(x))}{\lambda + l_u(x)} + \frac{1}{\lambda + l_u(x)} \sum_{y\neq x} L_{u(x)}(x,y) \cdot J_\lambda(y,u).
\end{aligned}
$$

In (a) we used the fact that $t_1$ is exponentially distributed with parameter $l_u(x)$, whereas (b) follows from the Markov property of the process and the properties of the generator. □

The right-hand side of equation (1.5) can be understood as the application of a linear operator on the cost functional $J_\lambda$:

Given an action $a \in \mathcal{A}$ and a real-valued function $J$ on $\mathcal{S}$, define the operator $T_a$ by

$$(T_aJ)(x) := \frac{c(x,a)}{\lambda + l_a(x)} + \frac{1}{\lambda + l_a(x)} \sum_{y\neq x} L_a(x,y)J(y). \tag{1.6}$$

In the case of infinite $\mathcal{S}$, we assume $J$ to be non-negative in order to guarantee that $T_a$ is well defined.
In line with this, we define the operator $T_u$ for a (deterministic stationary) policy $u \in \mathcal{U}$ by

$$(T_uJ)(x) := (T_{u(x)}J)(x) = \frac{c(x,u(x))}{\lambda + l_u(x)} + \frac{1}{\lambda + l_u(x)} \sum_{y\neq x} L_{u(x)}(x,y)J(y). \tag{1.7}$$

By this notation and $J_\lambda^u(x) := J_\lambda(x,u)$, equation (1.5) can be written as $J_\lambda^u(x) = (T_{u(x)}J_\lambda^u)(x)$ resp. $J_\lambda^u = T_uJ_\lambda^u$, which is a common representation [26]. This means that $J_\lambda^u$ is a fix-point of the operator $T_u$. The question is whether the fix-point equation

$$J = T_uJ. \tag{1.8}$$

uniquely characterizes the function $J_\lambda^u$. For a **finite state space** the answer to this question is yes.

**Lemma 1.5.** *For finite $\mathcal{S}$, the operator $T_u$ is a contraction on $(\mathbb{R}^{|\mathcal{S}|}, ||\cdot||_\infty)$ for all $u \in \mathcal{U}$.*

*Proof.* For $J, \tilde{J} \in \mathbb{R}^{|\mathcal{S}|}$ it holds

$$\begin{aligned}
||T_uJ - T_u\tilde{J}||_\infty &= \max_{x\in\mathcal{S}} |(T_uJ)(x) - (T_u\tilde{J})(x)| \\
&= \max_{x\in\mathcal{S}} \left| \frac{1}{\lambda + l_u(x)} \sum_{y\neq x} L_{u(x)}(x,y) \cdot \Big( J(y) - \tilde{J}(y) \Big) \right| \\
&\leq \max_{x\in\mathcal{S}} \frac{1}{\lambda + l_u(x)} \sum_{y\neq x} L_{u(x)}(x,y) \cdot |J(y) - \tilde{J}(y)| \\
&\leq \max_{x\in\mathcal{S}} \frac{1}{\lambda + l_u(x)} \sum_{y\neq x} L_{u(x)}(x,y) \cdot ||J - \tilde{J}||_\infty \\
&= ||J - \tilde{J}||_\infty \max_{x\in\mathcal{S}} \frac{\sum_{y\neq x} L_{u(x)}(x,y)}{\lambda + l_u(x)} \\
&= ||J - \tilde{J}||_\infty \max_{x\in\mathcal{S}} \frac{l_u(x)}{\lambda + l_u(x)} \\
&= ||J - \tilde{J}||_\infty \cdot \alpha
\end{aligned}$$

with $\alpha := \max_{x\in\mathcal{S}} \frac{l_u(x)}{\lambda + l_u(x)} < 1$. □

By the contraction property of the operator $T_u$ and the uniqueness of its fix-point, we can directly deduce

**Corollary 1.6.** *For finite $\mathcal{S}$, the function $J_\lambda^u$ of expected discounted costs given a policy $u \in \mathcal{U}$ is uniquely defined as the solution of the fix-point equation* (1.8)*. In this case, the matrix $\lambda I - L_u$ (with $I \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ being the identity matrix) is invertible and it holds*

$$J_\lambda^u = (\lambda I - L_u)^{-1} c_u,$$

*where $L_u(x,y) := L_{u(x)}(x,y)$ and $c_u(x) := c(x, u(x))$ for all $x, y \in \mathcal{S}$.*

*Proof.* In order to show that $\lambda I - L_u$ is invertible, we assume that the equation $(\lambda I - L_u)v = 0$ has a solution $v \neq 0$. Let $x^* := \arg\max_{x\in\mathcal{S}} |v(x)|$, such that $||v||_\infty = |v(x^*)|$. From $(\lambda I - L)v = 0$ it follows especially

$$\lambda v(x^*) = \sum_{y\neq x^*} L(x^*, y)(v(y) - v(x^*)).$$

Assuming $v(x^*) > 0$ we get $\lambda v(x^*) > 0$ but $v(y) \leq v(x^*)$ for all $y \neq x^*$ such that $\sum_{y\neq x^*} L(x,y)(v(y) - v(x^*)) \leq 0$. On the other hand, $v(x^*) \leq 0$ means $\lambda v^(x^*) \leq 0$ but $v(y) \geq v(x^*)$ for all $y \neq x^*$ such that $\sum_{y\neq x^*} L(x,y)(v(y) - v(x^*)) \geq 0$. In both cases this gives a contradiction. □

By Corollary 1.6, we can directly deduce a property of the operator $T_u$ which will be useful for the analysis of optimal policies as well as for the numerical approaches in Section 1.2.

**Lemma 1.7.** *If a non-negative function $J$ on $\mathcal{S}$ fulfills*

$$J(x) \geq (T_u J)(x) \quad \forall x \in \mathcal{S}, \tag{1.9}$$

*then $J(x) \geq J_\lambda(x, u)$ holds for all $x \in \mathcal{S}$.*

*Proof.* We give the proof for finite $\mathcal{S}$, making use of Corollary 1.6. For infinite $\mathcal{S}$ we refer to [26]. From (1.9) we can conclude that there exists a non-negative function $f$ on $\mathcal{S}$ with

$$\begin{aligned} J(x) &= (T_u J)(x) + \frac{f(x)}{\lambda + l_u(x)} \\ &= \frac{c(x, u(x)) + f(x)}{\lambda + l_u(x)} + \frac{1}{\lambda + l_u(x)} \sum_{y \neq x} L_{u(x)}(x, y) J(y) \\ &=: (\tilde{T}_u J)(x), \end{aligned}$$

where $\tilde{T}$ is an operator corresponding to the considered Markov control model with modified cost function $\tilde{c}(x, a) := c(x, a) + f(x)$. From Corollary 1.6 we know that $J$ has to be the cost functional $\tilde{J}_\lambda(x, u)$ for this modified model, i.e.

$$J(x) = \tilde{J}_\lambda(x, u) = \mathbb{E}_x^u \left( \int_0^\infty e^{-\lambda s} \tilde{c}(X_s, u(X_s))\, ds \right),$$

which is monotone in $\tilde{c}$, such that

$$J(x) \geq J_\lambda(x, u) \quad \forall x \in \mathcal{S}$$

because it holds $\tilde{c}(x, a) \geq c(x, a)$ for all $x \in \mathcal{S}$, $a \in \mathcal{A}$. □

In the case of infinite $\mathcal{S}$, the cost functional $J_\lambda^u$ is the *minimum non-negative solution* of the fix-point equation (1.8), see [26]. Instead of a convergence $||(T_u)^n J - J_\lambda^u||_\infty \overset{n\to\infty}{\longrightarrow} 0$ (uniform convergence), there is a pointwise convergence $(T_u^n J_0)(x) \to J_\lambda^u(x)\ \forall x \in \mathcal{S}$ for $J_0 = 0$.
Moreover, for any $\mathcal{S}$, the operator $T_u$ (resp. $T_a$) is obviously monotone, i.e. $\tilde{J}(x) \geq J(x)\ \forall x \in \mathcal{S}$ implies $(T_u \tilde{J})(x) \geq (T_u J)(x)\ \forall x \in \mathcal{S}$. Together this means:
By defining

$$J_0 := 0, \quad J_{n+1} := T_u J_n \quad \text{for } n \geq 0$$

we get a non-decreasing sequence $(J_n)_{n \in \mathbb{N}_0}$ with $J_n(x) \overset{n\to\infty}{\longrightarrow} J_\lambda(x, u)$ for all $x \in \mathcal{S}$ ($\mathcal{S}$ finite or infinite).

### Optimal policy and value function

So far we considered the cost functional $J_\lambda(x,u)$ for a given policy $u \in \mathcal{U}$ and discovered it to be the fix-point of the monotone operator $T_u$. Now, the essential problem is to characterize a policy which optimizes the costs and to calculate the corresponding optimal value of the cost functional. In general, the existence of an optimal policy cannot be taken for granted; here we will need further constraints. Before we turn to this issue, we give a characterization of the value function

$$V_\lambda(x) = \inf_{u\in\mathcal{U}} J_\lambda(x,u).$$

The corresponding results do not require the state space to be finite.

**Theorem 1.8** (Discounted cost optimality equation/Bellman equation)**.** *The value function $V_\lambda$ satisfies the recursion*

$$V_\lambda(x) = \inf_{a\in\mathcal{A}(x)} \left\{ \frac{c(x,a)}{\lambda + l_a(x)} + \frac{1}{\lambda + l_a(x)} \sum_{y\neq x} L_a(x,y)V_\lambda(y) \right\} \quad \forall x \in \mathcal{S}. \tag{1.10}$$

The Bellman equation (1.10), also known as *dynamic programming equation*, is named after its discoverer Richard Bellman. The dynamic programming concept consists of solving complex problems by breaking them down into simpler subproblems – which here refers to converting a minimization problem on the (huge) set of policies into a pointwise minimization for each state $x \in \mathcal{S}$. This is the essential step to enable an effective numerical calculation of the optimal policy.
Theorem 1.8 is a well-established result of Markov control theory, see e.g. [26]. We will formulate the proof by means of the operators $T_a$ resp. $T_u$ introduced in (1.6) resp. (1.7).

*Proof of Theorem 1.8.* Equation (1.10) can be written as $V_\lambda(x) = \inf_{a\in\mathcal{A}(x)}(T_aV_\lambda)(x)$. For an arbitrary policy $u \in \mathcal{U}$ it holds by definition that $J^u_\lambda(x) \geq V_\lambda(x)$ such that (as $T_a$ is a monotone operator for any $a \in \mathcal{A}$)

$$J^u_\lambda(x) = (T_{u(x)}J^u_\lambda)(x) \geq (T_{u(x)}V_\lambda)(x) \geq \inf_{a\in\mathcal{A}(x)}(T_aV_\lambda)(x).$$

This yields

$$V_\lambda(x) = \inf_{u\in\mathcal{U}} J^u_\lambda(x) \geq \inf_{a\in\mathcal{A}(x)}(T_aV_\lambda)(x).$$

On the other hand, assuming the existence of an $x^* \in \mathcal{S}$ with the property $V_\lambda(x^*) > \inf_{a\in\mathcal{A}(x^*)}(T_aV_\lambda)(x^*)$ yields

$$V_\lambda(x) = \inf_{a\in\mathcal{A}(x)}(T_aV_\lambda)(x) + \varepsilon(x) \quad \forall x \in \mathcal{S},$$

where $\varepsilon(x) \geq 0$ for all $x \in \mathcal{S}$ and $\varepsilon(x^*) > 0$. Depending on $x$ we can choose an action $a = u(x)$ such that

$$\begin{aligned} V_\lambda(x) &\geq (T_u V_\lambda)(x) + \frac{\varepsilon(x)}{2} \\ &= \frac{c(x, u(x)) + (\lambda + l_u(x))\frac{\varepsilon(x)}{2}}{\lambda + l_u(x)} + \frac{1}{\lambda + l_u(x)} \sum_{y \neq x} L_{u(x)}(x, y) V_\lambda(y) \\ &=: (\tilde{T}_u V_\lambda)(x), \end{aligned}$$

where $\tilde{T}_u$ is an operator referring to a new cost function $\tilde{c}(x, a) := c(x, a) + (\lambda + l_u(x))\frac{\varepsilon(x)}{2}$. With $\tilde{c}(x^*, a) > c(x^*, a)$ we have

$$\tilde{J}_\lambda(x^*, u) = \mathbb{E}^u_{x^*}\left(\int_0^\infty e^{-\lambda s} \tilde{c}(X_s, u(X_s))\, ds\right) > J_\lambda(x^*, u)$$

for the cost functional determined by $\tilde{c}$ and $\tilde{T}_u$. By Lemma 1.7 we can deduce $V_\lambda(x^*) \geq \tilde{J}_\lambda(x^*, u) > J_\lambda(x^*, u)$, in contradiction to $V_\lambda(x^*) = \inf_{u \in \mathcal{U}} J_\lambda(x^*, u)$. □

**Lemma 1.9.** *Assume $V_\lambda(x) < \infty$ for all $x \in \mathcal{S}$. Then the Bellman equation* (1.10) *is equivalent to*

$$\lambda V_\lambda(x) = \inf_{a \in \mathcal{A}(x)} \{c(x, a) + (L_a V_\lambda)(x)\} \quad \forall x \in \mathcal{S}.$$

For the proof of Lemma 1.9 we refer to [23].

The Bellman equation (1.10) characterizes not only the value function $V_\lambda$ but also the discounted-cost optimal policy, assuming that for each $x \in \mathcal{S}$ the infimum in (1.10) can be replaced by a minimum.

**Theorem 1.10.** *Suppose that there exists a policy $u^* \in \mathcal{U}$ which attains the minimum in the Bellman equation* (1.10)*, i.e.*

$$\begin{aligned} V_\lambda(x) &= \frac{c(x, u^*(x))}{\lambda + l_{u*}(x)} + \frac{1}{\lambda + l_{u*}(x)} \sum_{y \neq x} L_{u^*(x)}(x, y) V_\lambda(y) \qquad (1.11) \\ &= \min_{a \in \mathcal{A}(x)} \left\{ \frac{c(x, a)}{\lambda + l_a(x)} + \frac{1}{\lambda + l_a(x)} \sum_{y \neq x} L_a(x, y) V_\lambda(y) \right\} \quad \forall x \in \mathcal{S}. \end{aligned}$$

*Then $u^*$ is discounted-cost optimal, i.e. it holds $V_\lambda(x) = J_\lambda(x, u^*)$ for all $x \in \mathcal{S}$.*

*Proof.* Compare Theorem 4.10 in [26]: By definition it holds $V_\lambda(x) \leq J_\lambda(x, u^*)$. On the other hand, equation (1.11) means $V_\lambda = T_{u^*} V_\lambda$, such that by Lemma 1.7 we get $V_\lambda \geq J_\lambda(x, u^*)$. Together we get $V_\lambda = J_\lambda(x, u^*)$. □

The next manifesting question is under which conditions the existence of a policy $u^*$ fulfilling (1.11) can be guaranteed. It is common practice to make the following assumption.

**Assumption 1.11.**
a) The set of available actions $\mathcal{A}(x)$ is compact for each state $x \in \mathcal{S}$.
b) For all $x, y \in \mathcal{S}$, the functions $c(x,a)$ and $L_a(x,y)$ are continuous on $\mathcal{A}(x)$.
c) The cost function $c(x,a)$ is bounded on $\mathcal{S} \times \mathcal{A}$.

Note that for finite $\mathcal{S}$ part c) of Assumption 1.11 directly follows from a) and b). There exist several conditions that verify Assumption 1.11, see for example [24,27,48]. One possibility is to assume both the state space $\mathcal{S}$ and the action space $\mathcal{A}$ to be finite. Although this restraint is quite restrictive, we will choose it here because it allows for a compact proof of Theorem 1.12 and at the same time guarantees the convergence of the policy iteration algorithm which will be presented in Section 1.2. For the more general setting (and appropriate formulations of Theorem 1.12) we again refer to [26].

Let $\mathcal{S}$ be finite. In the style of [23, 26], we define the operator $T^* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ by

$$\begin{aligned}(T^*J)(x) \quad &:= \quad \inf_{a\in\mathcal{A}(x)} \left\{ \frac{c(x,a)}{\lambda + l_a(x)} + \frac{1}{\lambda + l_a(x)} \sum_{y\neq x} L_a(x,y)J(y) \right\} \\ &= \quad \inf_{a\in\mathcal{A}(x)} (T_aJ)(x).\end{aligned}$$

For finite $\mathcal{A}$ we replace “inf” by “min”.

**Theorem 1.12.** *Assume both $\mathcal{S}$ and $\mathcal{A}$ to be finite and define*

$$V_0 := 0, \quad V_{n+1} := T^*V_n \quad \textit{for}\, n \geq 0.$$

*For each $n \geq 0$, let $u_n \in \mathcal{U}$ be such that $V_{n+1}(x) = (T_{u_n(x)}V_n)(x)$, i.e. $u_n$ chooses the minimum argument in $(T^*V_n)(x) = \min_{a\in\mathcal{A}(x)}(T_aV_n)(x)$ depending on $x \in \mathcal{S}$. Then*

*a) $(V_n)_{n\in\mathbb{N}}$ is non-decreasing in $n \geq 0$.*

*b) $\lim_{n\to\infty} V_n(x) = V_\lambda(x)$ for all $x \in \mathcal{S}$.*

*c) There is a policy $u^* \in \mathcal{U}$ with $u_n(x) \overset{n\to\infty}{\longrightarrow} u^*(x)$ for all $x \in \mathcal{S}$, and $u^*$ is discounted-cost optimal, i.e. $J_\lambda(x,u^*) = V_\lambda(x)$ for all $x \in \mathcal{S}$.*

*Proof.* First of all, we observe that $V_n \leq J^u_\lambda$ for all $u \in \mathcal{U}$ and $n \in \mathbb{N}$, which follows by induction: Clearly $V_0 = 0 \leq J^u_\lambda$, and assuming $V_n \leq J^u_\lambda$ for an $n \in \mathbb{N}$ one can deduce $V_{n+1} = T^*V_n = \inf_u T_uV_n \leq T_uV_n \leq T_uJ^u_\lambda = J^u_\lambda$.
Now a) and b) immediately follow from the fact that all $T_u$ are contractions.
With $\mathcal{S}$ and $\mathcal{A}$ also the set $\mathcal{U}$ of policies is finite such that there exists a $u^* \in \mathcal{U}$ with $V_\lambda = \min_u J^u_\lambda = J^{u^*}_\lambda$. For $u \in \mathcal{U}$ denote by $\alpha_u := \max_{x\in\mathcal{S}} \frac{l_u(x)}{\lambda+l_u(x)}$ the contraction constant of $T_u$. Set $\varepsilon = \min_u \frac{1-\alpha_u}{\alpha_u}||J_u - V_\lambda||_\infty$. By b) there is an $n_0 \in \mathbb{N}$ such that

$0 < V_\lambda(x) - V_n(x) < \varepsilon$ for all $n \geq n_0$, $x \in \mathcal{S}$. Applying $T_u$, $u \neq u^*$, to $V_n$ for $n \geq n_0$ we have

$$\begin{aligned}
||J_u - T_u V_n||_\infty &= ||T_u J_u - T_u V_n||_\infty \\
&\leq \alpha_u ||J_u - V_n||_\infty \\
&\leq \alpha_u(||J_u - V_\lambda||_\infty + ||V_\lambda - V_n||_\infty) \\
&< \alpha_u(||J_u - V_\lambda||_\infty + \varepsilon) \\
&\leq \alpha_u(||J_u - V_\lambda||_\infty + \frac{1-\alpha_u}{\alpha_u}||J_u - V_\lambda||_\infty) \\
&= ||J_u - V_\lambda||_\infty .
\end{aligned}$$

Thus, for each $u \neq u^*$, there is $x \in \mathcal{S}$ with $|J_u(x) - V_\lambda(x)| = ||J_u - V_\lambda||_\infty > ||J_u - T_u V_n||_\infty \geq |J_u(x) - (T_u V_n)(x)|$. As $J_u \geq V_\lambda$, we get $V_{n+1}(x) = (T_u V_n)(x) > V_\lambda(x)$ which stands in contradiction to a) and b).
This means that for $n \geq n_0$ it has to be $u_n = u^*$, which completes the proof. □

**Example 1.2 (cont).** *We calculate the discounted costs for the 2-state-example 1.2 from page 9 and all possible policies, see Table 1.1. We set $\lambda = 0.1$ as well as $c_\mathcal{S}(x_2) = 10$ and $c_\mathcal{A}(a_2) = 2$, where $c(x,a) = c_\mathcal{S}(x) + c_\mathcal{A}(a)$. Remember that state $x_1$ and action $a_1$ are assumed to be cost-free.*

| $u(x_1)$ | $u(x_2)$ | $J_\lambda(x_1,u)$ | $J_\lambda(x_2,u)$ |
|---|---|---|---|
| $a_1$ | $a_1$ | 8.33 | 91.67 |
| $a_1$ | $a_2$ | 5.71 | 62.86 |
| $a_2$ | $a_1$ | 58.10 | 96.19 |
| $a_2$ | $a_2$ | 53.33 | 86.67 |

Table 1.1: **Discounted costs.** The expected discounted costs $J_\lambda(x,u)$ depending on the policy $u \in \mathcal{U}$ for the 2-state-example 1.2.

*We see that, as supposed, the optimal policy is given by $u^*(x_1) = a_1$, $u^*(x_2) = a_2$, and the value function fulfills*

$$V_\lambda(x_1) = 5.71 = 0 + \frac{0.01 \cdot 62.86}{0.1 + 0.01} = \frac{c(x_1,a_1)}{\lambda + l_1(x_1)} + \frac{L_1(x_1,x_2)V_\lambda(x_2)}{\lambda + l_1(x_1)}$$

*and*

$$V_\lambda(x_2) = 62.86 = \frac{12}{0.1+0.1} + \frac{0.1 \cdot 5.71}{0.1+0.1} = \frac{c(x_2,a_2)}{\lambda + l_2(x_2)} + \frac{L_2(x_2,x_1)V_\lambda(x_1)}{\lambda + l_2(x_2)}.$$

In Example 1.2 a straightforward comparison of all possible policies was feasible due to the small number of states and actions. In general, however, the number of policies will be very large making such a direct approach impossible. Instead, the optimal policy can be calculated by two types of dynamic programming algorithms that will be presented in Section 1.2.
Before we turn to the numerical realization, we now consider the average-cost criterion and deduce equivalent analytical results.

### 1.1.2 The Average-Cost Criterion

Recall that the average-cost criterion is given by

$$\bar{J}(x,u) = \limsup_{T\to\infty} \ \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T c(X_s, u(X_s))\, ds \right),$$

compare (1.3). Parallel to the case of discounted costs we tend to first characterize the cost functional $\bar{J}(x,u)$ for a given policy $u$ and then deduce some optimality criteria. There exists a lot of literature concerned with the average-cost criterion, see e.g. [25, 59, 81, 82]. All the approaches require several assumptions which are rather complicated than transparent. For instance, a typical restraint goes back to the discounted-cost criterion and assumes that there exists a state $x_0 \in \mathcal{S}$ and some decreasing sequence of discount factors with $\lambda_n \to 0$ such that $\lambda_n V_\lambda(x_0)$ is bounded in $n \in \mathbb{N}$.

Instead of following the common practice, we will give an own approach by restricting the analysis to finite state spaces and ergodic dynamics. More precisely, we assume that for each policy $u \in \mathcal{U}$, the dynamics of the controlled process are ergodic in the following sense: Given a policy $u \in \mathcal{U}$, there exists a unique equilibrium distribution $\mu_u$ on $\mathcal{S}$ fulfilling $\mu_u L_u = 0$ [56] (where as before $L_u(x,y) := L_{u(x)}(x,y)$, see Corollary 1.6) such that

$$\bar{J}(x,u) = \lim_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T c(X_s, u(X_s))\, ds \right) = \langle \mu_u, c_u \rangle \quad \forall x \in \mathcal{S},$$

where $c_u(x) := c(x, u(x))$ and $\langle \cdot, \cdot \rangle$ refers to the standard inner product on $\mathbb{R}^{|\mathcal{S}|}$. In this case, the function $\bar{J}(x,u)$ of long-run expected average costs is independent of $x \in \mathcal{S}$ and we can define the constant

$$\eta_u := \bar{J}(x,u)$$

of average costs given policy $u \in \mathcal{U}$.

#### Average costs $\eta_u$ for a given policy $u \in \mathcal{U}$

For the discounted-cost criterion we showed that the cost functional $J_\lambda^u$ of a given policy fulfills a recursion (see Lemma 1.4) and that for finite state spaces this recursion uniquely determines the cost functional which results in an analytic expression for $J_\lambda^u$ in terms of the given cost function and the generator $L_u$. Equivalent statements for the average-cost criterion and a finite state space will be given in the following lemma. Here, we formulate all parts together because the proofs are connected.

**Lemma 1.13.** *Let the state space $\mathcal{S}$ be finite.*

*a) Given a policy $u \in \mathcal{U}$, there exists a function $v : \mathcal{S} \to \mathbb{R}$ such that the corresponding constant $\eta_u$ of long-run expected average costs fulfills the equation*

$$\eta_u = c(x, u(x)) + \sum_{y\in\mathcal{S}} L_{u(x)}(x,y) v(y) \quad \forall x \in \mathcal{S}. \tag{1.12}$$

*b) The constant* $\eta_u$ *is uniquely determined by* (1.12) *and coincides with the first component of the vector*

$$(E - L_u)^{-1} c_u,$$

*where*

$$E := \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}.$$

*Proof.* We begin with part b) and show the uniqueness of the constant.
Assume that $\rho \in \mathbb{R}$ and $v : \mathcal{S} \to \mathbb{R}$ fulfill

$$\rho = c(x, u(x)) + \sum_{y \in \mathcal{S}} L_{u(x)}(x, y) v(y) \quad \text{for all } x \in \mathcal{S}. \tag{1.13}$$

Due to the structure of the generator matrix $L_u$ it holds $L_u(v + d) = L_u v$ for any constant vector $d \in \mathbb{R}^{|\mathcal{S}|}$ such that we can set $v(1) = \rho$ without loss of generality. Equation (1.13) can now be written as

$$Ev = c_u + L_u v$$

which yields

$$v = (E - L_u)^{-1} c_u.$$

The matrix $E - L_u$ is invertible by the following argumentation. If it was not invertible the equation

$$(E - L_u) w = 0 \tag{1.14}$$

would have a solution $w \neq 0$. Now, equation (1.14) is equivalent to $w(1) = (L_u w)(x)$ for all $x \in \mathcal{S}$. However, for $x_{\min} := \arg\min w(x)$ and $x_{\max} := \arg\max w(x)$ we have

$$(L_u w)(x_{\min}) = \sum_{y \neq x_{\min}} L_u(x_{\min}, y) \cdot (w(y) - w(x_{\min})) \geq 0$$

and

$$(L_u w)(x_{\max}) = \sum_{y \neq x_{\max}} L_u(x_{\max}, y) \cdot (w(y) - w(x_{\max})) \leq 0.$$

This leads to $0 \leq (L_u w)(x_{\min}) = w(1) = (L_u w)(x_{\max}) \leq 0$ which means that $w(1)$ has to be zero, and thus $L_u w = 0$ holds. This equation, however, is fulfilled by any constant vector $w$. As we assumed the process to be ergodic, the eigenvalue 0 is of multiplicity one, which means that such a constant $w$ is the only possible choice. This implies $w(x) = w(1) = 0$ for all $x$, in contradiction to $w \neq 0$.
This means that, given the side constraint $v(1) = \rho$, the quantities $\rho$ and $v$ are uniquely defined by $L_u$ and $c_u$. (Without this side constraint, the constant $\rho$ is still unique, whereas $v$ can be replaced by $v + d$ for any constant vector $d \in \mathbb{R}^{|\mathcal{S}|}$.)
It remains to show that $\rho = \eta_u$ which is part a) of the theorem. Fixing the generator $L_u$, we consider the function $f : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ given by $f(c) = \left((E - L_u)^{-1} c\right)(1) =$

$c(x) + (L_u(E - L_u)^{-1}c)(x) = c(x) + (L_u v_c)(x)$, independent of $x \in \mathcal{S}$. (We write $v_c$ in order to underline the dependence of the vector $v$ on the cost function $c$.) This function delivers the constant $\rho$ depending on the cost function $c$. As $f$ is obviously linear in $c$, by the Riesz representation theorem (see e.g. [71]) there exists a vector $w \in \mathbb{R}^{|\mathcal{S}|}$ such that

$$f(c) = \langle w, c \rangle.$$

Applying $f$ to a vector of the form $L_u c$ yields $f(L_u c) = (L_u c)(x) + (L_u(E - L_u)^{-1}c)(x) = (L_u c)(x) + (L_u v_c)(x) = L_u(c + v_c)(x) = \rho$ for all $x$. Now the last equality corresponds to equation (1.13) by setting the cost function to zero such that the constant $\rho$ is given by $f(0) = \langle w, 0 \rangle = 0$. We get $f(L_u c) = 0$ and with it

$$f(L_u c) = \langle w, L_u c \rangle = \langle L_u' w, c \rangle = 0$$

for all $c \in \mathbb{R}^{|\mathcal{S}|}$. By choosing $c(1) = 1$ and $c(x) = 0$ for all $x \neq 1$ we get $\langle L_u' w, c \rangle = (L_u' w)(1) = 0$, and equivalently we can deduce $(L_u' w)(x) = 0$ for all other $x \in \mathcal{S}$. However, the resulting equation $L_u' w = 0$ is exactly the characterization for the equilibrium distribution $\mu_u$ such that $w = \mu_u$ and with it

$$f(c) = \langle \mu_u, c \rangle = \eta_u$$

which proves a). □

Part a) of Lemma 1.13 is also true for non-finite state spaces; however, as mentioned above, this requires several additional assumptions (e.g. that the jump rates are uniquely bounded on $\mathcal{S}$, i.e. $\sup_{x \in \mathcal{S}} \sup_{a \in \mathcal{A}(x)} l_a(x) < \infty$), see [25, 81] among others.
In analogy to Lemma 1.7 of the discounted-cost criterion, we now formulate the following monotony property which holds for arbitrary (denumerable) $\mathcal{S}$.

**Lemma 1.14.** *Let $u \in \mathcal{U}$ be a given policy. It holds:*

*a) If there exists a constant $g \geq 0$ and a non-negative function $v$ on $\mathcal{S}$ such that*

$$g \geq c(x, u(x)) + \sum_{y \in \mathcal{S}} L_{u(x)}(x, y) v(y) \quad \forall x \in \mathcal{S}, \tag{1.15}$$

*then $g \geq \eta_u$.*

*b) If there exists a constant $g \geq 0$ and a non-negative function $v$ on $\mathcal{S}$ such that*

$$g \leq c(x, u(x)) + \sum_{y \in \mathcal{S}} L_{u(x)}(x, y) v(y) \quad \forall x \in \mathcal{S}, \tag{1.16}$$

*then $g \leq \eta_u$.*

*Proof.* We make use of Lemma 1.13 to prove both statements for finite $\mathcal{S}$. For infinite $\mathcal{S}$ we refer to [26].

a) Define

$$h(x) := g - \left( c(x, u(x)) + \sum_{y \in \mathcal{S}} L_{u(x)}(x, y) v(y) \right)$$

such that

$$g = h(x) + c(x, u(x)) + \sum_{y \in \mathcal{S}} L_{u(x)}(x, y) v(y) \quad \forall x \in \mathcal{S}.$$

By (1.15) we have $h(x) \geq 0$ for all $x \in \mathcal{S}$. As $\tilde{c}(x, u(x)) := h(x) + c(x, u(x)) \geq c(x, u(x))$ and

$$g = \tilde{c}(x, u(x)) + \sum_{y \in \mathcal{S}} L_{u(x)}(x, y) v(y) \quad \forall x \in \mathcal{S}$$

we can infer from Lemma 1.13 that $g$ is the constant of average costs under the given policy $u$ for the cost function $\tilde{c}$, i.e.

$$\begin{aligned} g &= \lim_{T \to \infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T \tilde{c}(X_s, u(X_s)) \, ds \right) \\ &\geq \lim_{T \to \infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T c(X_s, u(X_s)) \, ds \right) \\ &= \eta_u, \end{aligned}$$

which completes the proof of a).
Part b) is analogous. □

**Optimal policy and value function**

Being aware that the constant $\eta_u$ of long-term average costs for a policy $u \in \mathcal{U}$ is part of the solution of a system of linear equations, compare Lemma 1.13, we are now interested in the characterization of the optimal average costs

$$\eta^* := \inf_{u \in \mathcal{U}} \eta_u.$$

To this end, we will again formulate an optimality equation which forms the background for the numerical calculation of an optimal policy in Section 1.2. Interestingly, the proof will resort to the results for the discounted-cost problem from Section 1.1.1. Moreover, we will make use of Lemma 1.13 and Lemma 1.14 which both were formulated only for finite $\mathcal{S}$. For this reason, we again assume $\mathcal{S}$ to be finite for the following statements. However, all these statements are also true for non-finite $\mathcal{S}$, and even the proofs in principle coincide; we refer to [26].

In contrast to the case of discounted costs, the formulation of an average-cost optimality equation requires additional assumptions concerning the cost function, the generators and the sets $\mathcal{A}(x)$ of available actions. One possibility is to employ assumption 1.11, what we will do here because it allows for a compact analysis and suffices to get an insight into the structure of the control problem. For more general results we again refer to common Markov control literature.

**Theorem 1.15** (Average cost optimality equation/Bellman equation)**.** *Suppose that Assumption 1.11 holds. Then there exists a function* $v^* : \mathcal{S} \to \mathbb{R}$ *and a constant* $g \geq 0$ *satisfying*

$$g = \inf_{a \in \mathcal{A}(x)} \left\{ c(x,a) + \sum_{y \in \mathcal{S}} L_a(x,y) v^*(y) \right\} \quad \forall x \in \mathcal{S}. \tag{1.17}$$

*The constant* $g$ *coincides with the optimal average costs* $\eta^*$*.*

Note that the function $v^*$ is fixed and not part of the minimization argument.

*Proof.* The approach is taken from the proof of Theorem 5.7 in [26].
By Assumption 1.11 c) there exists a constant $c_{\max} > 0$ such that $0 \leq c(x,a) \leq c_{\max}$ for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$. From this we can deduce $0 \leq \lambda V_\lambda(x) \leq c_{\max}$ for all $x \in \mathcal{S}$ and $\lambda > 0$, where $V_\lambda$ is the value function of the corresponding discounted-cost problem. Moreover, fixing a state $x_0 \in \mathcal{S}$ and setting $h_\lambda(x) := V_\lambda(x) - V_\lambda(x_0)$, it holds $0 \leq |h_\lambda(x)| \leq c_{\max}$ for all $x \in \mathcal{S}$ and $\lambda > 0$.
The Bolzano-Weierstrass theorem gives the existence of a subsequence $(\lambda_n)_{n \in \mathbb{N}}$ of discount factors with $\lambda_n \overset{n\to\infty}{\longrightarrow} 0$, a constant $g$ and a function $v^*$ on $\mathcal{S}$ such that

$$g = \lim_{n\to\infty} \lambda_n V_{\lambda_n}(x_0) \quad \text{and} \quad v^*(x) = \lim_{n\to\infty} h_{\lambda_n}(x) \quad \forall x \in \mathcal{S}.$$

We note that it holds $\lambda_n h_{\lambda_n}(x) \overset{n\to\infty}{\longrightarrow} 0$ for all $x \in \mathcal{S}$ and thereby

$$\lambda_n V_{\lambda_n}(x) = \lambda_n h_{\lambda_n}(x) + \lambda_n V_{\lambda_n}(x_0) \overset{n\to\infty}{\longrightarrow} g \quad \forall x \in \mathcal{S}.$$

Now, for $\varepsilon > 0$ and $n \in \mathbb{N}$, Lemma 1.9 provides a policy $u_n \in \mathcal{U}$ with

$$\lambda_n V_{\lambda_n}(x) \geq c(x, u_n(x)) + (L_{u_n} V_{\lambda_n})(x) - \lambda_n \varepsilon \quad \forall x \in \mathcal{S}.$$

By part a) of Assumption 1.11 there exists a policy $u_0 \in \mathcal{U}$ and a subsequence $(u_k)_{k \in \mathbb{N}}$ of $(u_n)_{n \in \mathbb{N}}$ with $u_k(x) \overset{k\to\infty}{\longrightarrow} u_0(x)$ for all $x \in \mathcal{S}$ and, by part b) of Assumption 1.11,

$$\lim_{k\to\infty} c(x, u_k(x)) = c(x, u_0(x)), \quad \lim_{k\to\infty} L_{u_k}(x,y) = L_{u_0}(x,y) \quad \forall x, y \in \mathcal{S}.$$

Together we get

$$\begin{aligned} g &= \lim_{k\to\infty} \lambda_k V_{\lambda_k}(x) \\ &\geq \lim_{k\to\infty} \left( c(x, u_k(x)) + (L_{u_k} V_{\lambda_k})(x) - \varepsilon \lambda_k \right) \\ &= \lim_{k\to\infty} \left( c(x, u_k(x)) + (L_{u_k} h_{\lambda_k})(x) - \varepsilon \lambda_k \right) \\ &= c(x, u_0(x)) + (L_{u_0} v^*)(x) \\ &\geq \inf_{a \in \mathcal{A}(x)} \{ c(x,a) + (L_a v^*)(x) \} \quad \forall x \in \mathcal{S}. \end{aligned}$$

At the same time, using Lemma 1.9, we obtain

$$\lambda_n V_{\lambda_n}(x) \leq c(x,a) + (L_a V_{\lambda_n})(x) \quad \forall x \in \mathcal{S}, a \in \mathcal{A}(x),$$

and with it

$$\lambda_n V_{\lambda_n}(x) \leq c(x,a) + (L_a h_{\lambda_n})(x) \quad \forall x \in \mathcal{S}, a \in \mathcal{A}(x).$$

Taking the limit $n \to \infty$ on both sides delivers

$$g \leq c(x,a) + (L_a v^*)(x) \quad \forall x \in \mathcal{S}, a \in \mathcal{A}(x),$$

thus

$$g \leq \inf_{a \in \mathcal{A}(x)} \{c(x,a) + (L_a v^*)(x)\} \quad \forall x \in \mathcal{S}.$$

It remains to argue that $g = \eta^*$: On the one hand, by $g \geq c(x, u_0(x)) + (L_{u_0} h)(x)$ and Lemma 1.14 it holds $g \geq \eta_{u_0}$ and therefore $g \geq \eta^*$. On the other hand, for any policy $u \in \mathcal{U}$, it is

$$\begin{aligned} g &= \lim_{k\to\infty} \lambda_k V_{\lambda_k}(x) \\ &\leq \limsup_{k\to\infty} \; \lambda_k J_{\lambda_k}(x,u) \\ &\overset{(*)}{\leq} \limsup_{T\to\infty} \; \frac{1}{T}\mathbb{E}_x^u \left( \frac{1}{T} \int_0^T c(X_s, u(X_s))\, ds \right) \\ &= \eta_u, \end{aligned}$$

where $(*)$ goes back to the Tauberian theorem [26, 76]. This implies $g \leq \inf_u \eta_u = \eta_u^*$. □

As in the case of discounted costs, the Bellman equation (1.17) characterizes not only the optimal average costs but also the optimal policy (if existent), see e.g. [81].

**Theorem 1.16.** *Suppose that there exists a policy $u^* \in \mathcal{U}$ which attains the minimum in the Bellman equation* (1.17)*, i.e.*

$$\begin{aligned} g &= c(x, u^*(x)) + \sum_{y \in \mathcal{S}} L_{u^*(x)}(x,y) v^*(y) \\ &= \min_{a \in \mathcal{A}(x)} \left\{ c(x,a) + \sum_{y\in\mathcal{S}} L_a(x,y) v^*(y) \right\} \quad \forall x \in \mathcal{S}. \end{aligned} \tag{1.18}$$

*Then $u^*$ is average-cost optimal and it holds $\eta^* = \bar{J}(x, u^*) = g$ for all $x \in \mathcal{S}$.*

*Proof.* This obviously follows from Lemma 1.13 and Theorem 1.15. □

Again, Assumption 1.11 guarantees the existence of an optimal policy, and especially for finite $\mathcal{S}$ and $\mathcal{A}$ its existence is out of question.

**Example 1.2 (cont).** *We consider again the 2-state-example 1.2 and calculate the average costs depending on the policy. The results are given in Table 1.2. As for the discounted-cost criterion, the optimal policy is given by* $u^*(x_1) = a_1$*,* $u^*(x_2) = a_2$*.*

| $u(x_1)$ | $u(x_2)$ | $\eta$ |
|---|---|---|
| $a_1$ | $a_1$ | 5.00 |
| $a_1$ | $a_2$ | 1.09 |
| $a_2$ | $a_1$ | 9.27 |
| $a_2$ | $a_2$ | 7.00 |

Table 1.2: **Averaged costs.** The average costs depending on the policy $u \in \mathcal{U}$ for the 2-state-example 1.2.

## 1.2 Numerical Realization

In the previous chapter we have seen that – given a Markov control model and an optimization criterion – the optimal policy can be characterized by a Bellman equation. We will now analyze how these results can be used in order to determine the optimal policy numerically. There exist two kinds of dynamic programming algorithms, the so called *value iteration* and *policy iteration* approach. They will be presented here for both the discounted- and the average-cost criterion.
The value iteration approach introduced by Bellman (1957) [6] consists of a successive approximation of the value function. Given this approximation, the corresponding (optimal) policy can be determined. In contrast, a policy iteration runs through the set of policies and alternates between a *policy evaluation* step in which the value of the current policy is calculated and a *policy improvement* step in which the current policy is improved. This procedure was proposed by Howard (1960) [32].
The algorithmic complexity of policy iteration is a matter of actual research. For a fixed discount factor the policy iteration algorithm is polynomial in $|\mathcal{S}|$ and $|\mathcal{A}|$ [79]. The same is true for value iteration [43]. The dependence of the runtime on the discount factor, however, is not polynomial [31]. For average costs, policy iteration has exponential runtime [20]. Generally, policy iteration converges no more slowly than value iteration [48].

We will in the following **assume both $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{A}}$ to be finite** and, as in Section 1.1.2, consider the controlled process to be ergodic for all possible policies.

### Discounted costs

First of all, we consider the discounted-cost criterion $J_\lambda(x, u)$, compare (1.2). Here, Theorem 1.12 directly delivers the following value iteration scheme.

**The discounted-cost value iteration algorithm:**

1. Let $V_0 := 0$ and $k = 0$, and define a threshold $\varepsilon > 0$.

2. Obtain

$$V_{k+1}(x) = (T^* V_k)(x) = \min_{a \in \mathcal{A}(x)} \left\{ \frac{c(x,a)}{\lambda + l_a(x)} + \frac{1}{\lambda + l_a(x)} \sum_{y \neq x} L_a(x,y) V_k(y) \right\}$$

   and choose

$$u_{k+1}(x) = \arg \min_{a \in \mathcal{A}(x)} \left\{ \frac{c(x,a)}{\lambda + l_a(x)} + \frac{1}{\lambda + l_a(x)} \sum_{y \neq x} L_a(x,y) V_k(y) \right\}$$

   (such that $V_{k+1} = T_{u_{k+1}} V_k$).

3. If $||V_{k+1} - V_k||_\infty \leq \varepsilon$, then stop. Otherwise increment $k$ by 1 and return to Step 2.

The value function of optimal costs $V_\lambda$ is approximated by $V_k$ and the optimal policy is approximated by $u_k$. The advantage of the value iteration algorithm is the straightforward calculation: There is no need to take the inverse of a (possibly huge) matrix or to solve a system of linear equations, as it will be the case in the evaluation step of policy iteration. However, it can be observed that within the value iteration the policies $(u_k)_{k=0,1,2,\ldots}$ often become (exactly) optimal long before the sequence $(V_k)_{k=0,1,2,\ldots}$ of approximations converged, see [47] and Table 1.3 of the example that will follow. This motivates to relocate the iteration steps directly onto the set of policies, as it is done in policy iteration, compare [26].

**The discounted-cost policy iteration algorithm:**

1. Pick an arbitrary $u \in \mathcal{U}$. Let $k = 0$ and $u_0 := u$.

2. (Policy evaluation) Obtain $J_\lambda^k = (\lambda I - L_{u_k})^{-1} c_{u_k}$.

3. (Policy improvement) For each $x \in \mathcal{S}$, calculate

$$m(x) := \min_{a \in \mathcal{A}(x)} \left\{ c(x,a) + \sum_{y \in \mathcal{S}} L_a(x,y) J_\lambda^k(y) \right\}$$

   and set $a_k(x) := \arg\min_{a \in \mathcal{A}(x)} \left\{ c(x,a) + \sum_{y \in \mathcal{S}} L_a(x,y) J_\lambda^k(y) \right\}$.
   Define $u_{k+1}$ as follows:

$$u_{k+1}(x) := \begin{cases} a_k(x), & \text{if } m(x) < \lambda J_\lambda^k(x), \\ u_k(x), & \text{if } m(x) = \lambda J_\lambda^k(x). \end{cases}$$

4. If $u_{k+1} = u_k$, then stop. Otherwise increment $k$ by 1 and return to Step 2.

**Theorem 1.17.** *The discounted-cost policy iteration algorithm yields a discounted-cost optimal policy in a finite number of iterations.*

*Proof.* Let $(u_k)_{k=0,1,2,\ldots}$ be the sequence of policies obtained by the algorithm. By definition of $u_{k+1}$ it holds $\lambda J_\lambda^k(x) > c(x, u_{k+1}(x)) + \sum_{y\in\mathcal{S}} L_{u_{k+1}(x)}(x,y)J_\lambda^k(y)$ which is equivalent to $J_\lambda^k > T_{u_{k+1}}J_\lambda^k$. By Lemma 1.7 we get $J_\lambda^{k+1} \leq J_\lambda^k$ meaning that the cost functional decreases with increasing $k$. As the policies $u_k$, $k = 0,1,2,\ldots$, are all different and the number of policies is finite, the iterations must stop after a finite number $K \in \mathbb{N}$ of iterations. The function $J_\lambda^K(x) = J_\lambda(x, u_K)$ fulfills the Bellman equation (1.10), such that $u_K$ is discounted-cost optimal. □

The most "tricky" part of the policy iteration algorithm is the evaluation step: Here the inverse of the matrix $\lambda I - L_{u_k} \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ has to be calculated. For a small state space, however, this will pose no problem and the policy iteration usually performs better than the value iteration.
In order to get an impression of how the two algorithms work, we perform the calculations for the simple 2-state-example considered in the previous sections.

**Example 1.2 (cont).** *Reconsider the 2-state-example 1.2 with*

$$L_1 = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix}, \quad L_2 = \begin{pmatrix} -0.1 & 0.1 \\ 0.1 & -0.1 \end{pmatrix}$$

*and $c_\mathcal{S}(x_2) = 10$ and $c_\mathcal{A}(a_2) = 2$ combined with the discounted-cost criterion. Table 1.3 shows the steps for the value iteration algorithm choosing $\varepsilon = 0.001$, while Table 1.4 shows the results for the policy iteration algorithm. The value function is given by $V_\lambda(x_1) = 5.71$ and $V_\lambda(x_2) = 62.86$.*

| $k$ | $V_k(x_1)$ | $V_k(x_2)$ | $u_k(x_1)$ | $u_k(x_2)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | - | - |
| 1 | 0 | 60 | $a_1$ | $a_2$ |
| 2 | 5.45 | 60 | $a_1$ | $a_2$ |
| 3 | 5.45 | 62.73 | $a_1$ | $a_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 9 | 5.71 | 62.86 | $a_1$ | $a_2$ |

Table 1.3: **Value iteration for discounted costs.** The steps of the value iteration algorithm for the 2-state-example 1.2 given the discounted-cost criterion.

*We can observe that both algorithms deliver the same result (as they should do), however the value iteration needs more iteration steps than the policy iteration. The reason for this is the slow convergence of the value iteration function $V_k$: Although $u_1$ already coincides with the optimal policy, the corresponding value iteration function $V_1$ strongly differs from the value function $V_\lambda$.*

| $k$ | $V_k(x_1)$ | $V_k(x_2)$ | $u_k(x_1)$ | $u_k(x_2)$ |
|---|---|---|---|---|
| 0 | 8.33 | 91.67 | $a_1$ | $a_1$ |
| 1 | 5.71 | 62.86 | $a_1$ | $a_2$ |

Table 1.4: **Policy iteration for discounted costs.** The steps of the policy iteration algorithm for the 2-state-example 1.2 given the discounted-cost criterion. In fact, for any given initial policy the algorithm converges after one iteration step.

### Average costs

When considering the average-cost criterion in the setting of continuous-time MDPs on a discrete state space, it seems to be common to deal only with policy iteration. While for discrete-time MDPs there exist several value iteration approaches (see e.g. [2]), in the continuous-time case value iteration is usually not mentioned in the literature. In the following, we will present our own approach for a value iteration in the setting of continuous-time MDPs. Using the results of [60], we will translate the given continuous-time MDP into an equivalent discrete-time MDP and then apply an appropriate discrete-time value iteration algorithm.

Given the Markov control model $\big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{L_a : a \in \mathcal{A}\},\, c\big)$, we define for each $a \in \mathcal{A}$ a transition matrix $P_a \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ by

$$P_a(x,y) := \begin{cases} \frac{L_a(x,y)}{l_{\max}} & \text{for } x \neq y, \\ 1 - \frac{l_a(x)}{l_{\max}} & \text{for } x = y, \end{cases}$$

and a cost function $\tilde{c} : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ by

$$\tilde{c}(x,a) := \frac{c(x,a)}{l_{\max}},$$

where $l_{\max} := \max_{x\in\mathcal{S},a\in\mathcal{A}} l_a(x)$. Remember that $\mathcal{S}$ and $\mathcal{A}$ are assumed to be finite. We consider the discrete-time Markov control model $\big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{P_a : a \in \mathcal{A}\},\, \tilde{c}\big)$, as well as the corresponding Markov Decision Process $(X_n)_{n\in\mathbb{N}_0}$ and the average-cost criterion $\tilde{J}(x,u)$, see Remark 1.3. Serfozo [60] states that, for a given policy $u \in \mathcal{U}$, it holds

$$\bar{J}(x,u) = l_{\max} \cdot \tilde{J}(x,u),$$

resp.

$$\eta_u = l_{\max} \cdot \tilde{\eta}_u,$$

where $\bar{J}(x,u)$ resp. $\eta_u$ are the average costs of the continuous-time MDP we started with, see Section 1.1. This means that the optimal policy of the discrete setting coincides with the optimal policy of the continuous setting.
The Bellman equation for the discrete setting is given by

$$\tilde{\eta}^* + v(x) = \min_{a\in\mathcal{A}(x)} \left\{ \tilde{c}(x,a) + \sum_{y\in\mathcal{S}} P_a(x,y)v(y) \right\},$$

where $v : \mathcal{S} \to \mathbb{R}$ [7].

For the discrete-time MDP we now consider the so called *relative value iteration* algorithm which is due to White [73], see also [2]:

1. Let $V_0 := 0$, $\rho_0 := 0$ and $k = 0$. Define a threshold $\varepsilon > 0$ and pick an arbitrary $x_0 \in \mathcal{S}$.[1]

2. Obtain

$$\rho_{k+1} = \min_{a \in \mathcal{A}(x)} \left\{ \tilde{c}(x_0, a) + \sum_{y \in \mathcal{S}} P_a(x_0, y) V_k(y) \right\}$$

and

$$V_{k+1}(x) = \min_{a \in \mathcal{A}(x)} \left\{ \tilde{c}(x, a) + \sum_{y \in \mathcal{S}} P_a(x, y) V_k(y) \right\} - \rho_{k+1}$$

and choose

$$u_{k+1}(x) = \arg \min_{a \in \mathcal{A}(x)} \left\{ \tilde{c}(x, a) + \sum_{y \in \mathcal{S}} P_a(x, y) V_k(y) \right\}.$$

3. If $k > 1$ and $|\rho_{k+1} - \rho_k| \leq \varepsilon$, then stop. Otherwise increment $k$ by 1 and return to Step 2.

The value $\tilde{\eta}^* = \min_{u \in \mathcal{U}} \tilde{\eta}_u$ of optimal average costs is approximated by $\rho_k$ and the corresponding $\varepsilon$-optimal policy is given by $u_k$, see Proposition 6 in Chapter 8.2 of [7]. In [59] one can find conditions for the convergence of the value iteration algorithm.
The algorithm is called "relative" because the values $V_k(x)$ are put in relation to $V_k(x_0)$ which avoids an unnecessary growth of $V_k(x)$ in $k$ and (in our calculations) results in a faster convergence.
In the notation of the continuous-time MDP this yields:

**The average-cost (relative) value iteration algorithm:**

1. Let $V_0 := 0$, $\rho_0 := 0$ and $k = 0$. Define a threshold $\varepsilon > 0$ and pick an arbitrary $x_0 \in \mathcal{S}$.

2. Obtain

$$\rho_{k+1} = \min_{a \in \mathcal{A}(x)} \left\{ \frac{c(x_0, a)}{l_{\max}} + V_k(x_0) + \sum_{y \in \mathcal{S}} \frac{L_a(x_0, y)}{l_{\max}} V_k(y) \right\}$$

[1] In [7] the state $x_0$ has to fulfill the following condition: There exist $\alpha > 0$ and $m \in \mathbb{N}$ such that $P_u^m(y, x_0) \geq \alpha$ for all $y \in \mathcal{S}$, $u \in \mathcal{U}$. As we assumed the state space to be finite and the process to be ergodic for all policies this assumption is naturally fulfilled for all states.

and

$$V_{k+1}(x) = \min_{a \in \mathcal{A}(x)} \left\{ \frac{c(x,a)}{l_{\max}} + V_k(x) + \sum_{y \in \mathcal{S}} \frac{L_a(x,y)}{l_{\max}} V_k(y) \right\} - \rho_{k+1}$$

and choose

$$u_{k+1}(x) = \arg \min_{a \in \mathcal{A}(x)} \left\{ \frac{c(x,a)}{l_{\max}} + V_k(x) + \sum_{y \in \mathcal{S}} \frac{L_a(x,y)}{l_{\max}} V_k(y) \right\}.$$

3. If $k > 1$ and $|\rho_{k+1} - \rho_k| \leq \varepsilon$, then stop. Otherwise increment $k$ by 1 and return to Step 2.

Here, $\rho_k \cdot l_{\max}$ delivers an approximation for $\eta^*$, and the final $u_k$ is (approximately) optimal.

The properties of the value iteration scheme for average costs are the same as for discounted costs: The calculations within an iteration step are comparatively simple, but the number of iteration steps can be very large. Again, the opposite is true for the following policy iteration scheme (derived from [82]).

**The average-cost policy iteration algorithm:**

1. Pick an arbitrary $u \in \mathcal{U}$. Let $k = 0$ and $u_k := u$.

2. (Policy evaluation) Find a constant $\eta_k$ and a real-valued function $v_k$ on $\mathcal{S}$ satisfying
$$\eta_k = c(x, u_k(x)) + \sum_{y \in \mathcal{S}} L_{u_k(x)}(x,y) v_k(y) \quad \forall x \in \mathcal{S}.$$

3. (Policy improvement) For each $x \in \mathcal{S}$ calculate
$$m(x) := \min_{a \in \mathcal{A}(x)} \left\{ c(x,a) + \sum_{y \in \mathcal{S}} L_a(x,y) v_k(y) \right\}$$
and set $a_k(x) := \arg\min_{a \in \mathcal{A}(x)} \left\{ c(x,a) + \sum_{y \in \mathcal{S}} L_a(x,y) v_k(y) \right\}$.
Define $u_{k+1}$ as follows:
$$u_{k+1}(x) := \begin{cases} a_k(x), & \text{if } m(x) < \eta_k, \\ u_k(x), & \text{if } m(x) = \eta_k. \end{cases}$$

4. If $u_{k+1} = u_k$, then stop. Otherwise increment $k$ by 1 and return to Step 2.

Parallel to Theorem 1.17 we can state that the average-cost policy iteration algorithm delivers an average-cost optimal policy in a finite number of iteration steps. The proof is analogous, making use of Lemma 1.14 and Theorem 1.15.

**Example 1.2 (cont).** *We apply both average-cost algorithms to the 2-state-example 1.2. Table 1.5 shows the steps for the value iteration algorithm choosing* $\varepsilon = 0.001$ *and* $x_0 = 1$*, while Table 1.6 shows the results for the policy iteration algorithm.*

*We can make the same observations as in the case of discounted costs: Both algorithms deliver the same result (the functions* $V_k$ *are equivalent because they differ only by a constant). Though, the value iteration requires more iteration steps than the policy iteration algorithm.*

| $k$ | $\rho_k$ | $V_k(x_1)$ | $V_k(x_2)$ | $u_k(x_1)$ | $u_k(x_2)$ |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | - | - |
| 1 | 0 | 0 | 100 | $a_1$ | $a_1$ |
| 2 | 10 | 0 | 110 | $a_1$ | $a_2$ |
| 3 | 11 | 0 | 109 | $a_1$ | $a_2$ |
| 4 | 10.9 | 0 | 109.1 | $a_1$ | $a_2$ |
| 5 | 10.91 | 0 | 109.09 | $a_1$ | $a_2$ |
| 6 | 10.909 | 0 | 109.091 | $a_1$ | $a_2$ |

Table 1.5: **Value iteration for average costs.** The steps of the relative value iteration algorithm for the 2-state-example 1.2 given the average-cost criterion. By $l_{\max} = 0.1$ we get $\rho_6 \cdot l_{\max} = 1.0909 = \eta^*$, compare Table 1.2.

| $k$ | $\eta_k$ | $V_k(x_1)$ | $V_k(x_2)$ | $u_k(x_1)$ | $u_k(x_2)$ |
|---|---|---|---|---|---|
| 0 | 5 | 5 | 505 | $a_1$ | $a_1$ |
| 1 | 1.0909 | 1.0909 | 110.1818 | $a_1$ | $a_2$ |

Table 1.6: **Policy iteration for average costs.** The steps of the policy iteration algorithm for the 2-state-example 1.2 given the average-cost criterion. For any given initial policy the algorithm converges after one iteration step.

Evidently, the considered example has very low dimensionality such that the optimal policy could have also been found by a direct comparison of the $|\mathcal{A}|^{|\mathcal{S}|} = 4$ policies. In the following example with $|\mathcal{A}|^{|\mathcal{S}|} = 2^{100}$ this is not the case.

**Example 1.18** (Controlled Population)**.** *We consider a simple birth-death-process of an "undesirable" population such as a virus or some harmful bacteria. The birth and death rates are denoted by* $\alpha$ *and* $\beta$*, respectively (*$\alpha, \beta \geq 0$*), and describe the rates at which the population increases or decreases in time [42]. The damage caused by the population is measured by some cost function* $c_{\mathcal{S}}(x)$ *which is assumed to be nondecreasing in the population size* $x \in \mathbb{N}_0$*. The process can be influenced by changing the birth and death rates.*

*More precisely, we are concerned with the state space* $\mathcal{S} = \mathbb{N}_0$ *and an action space* $\mathcal{A}$ *where each action* $a \in \mathcal{A}$ *is determined by the parameters* $\alpha_a$ *and* $\beta_a$*. The generators*

*$L_a$ have the form*

$$L_a = \begin{pmatrix} 0 & 0 & 0 & \cdots & & & \\ \beta_a & -(\beta_a+\alpha_a) & \alpha_a & 0 & \cdots & & \\ 0 & 2\beta_a & -2(\beta_a+\alpha_a) & 2\alpha_a & 0 & \cdots & \\ 0 & 0 & 3\beta_a & -3(\beta_a+\alpha_a) & 3\alpha_a & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \end{pmatrix}.$$

*An action $a \in \mathcal{A}$ produces costs at rate $c_{\mathcal{A}}(a)$, and the cost function $c$ is of the form $c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$.*

*In the following, we will present some numerical results for linear costs, $c_{\mathcal{S}}(x) = x$. We assume that there are only two actions $a_1$, $a_2$ with $\mathcal{A}(x) = \mathcal{A} = \{a_1, a_2\}$ for all $x \in \mathcal{S}$ and choose the discounted-cost criterion.*[2] *The infinite state space $\mathbb{N}_0$ is approximated by a finite one, $\mathcal{S} = \{0, ..., N\}$ with $N \in \mathbb{N}$ big enough. The generators and the cost function for this finite model are calculated via an algorithm deduced from [39].*

*We set $N = 100$ and $\lambda = 0.1$, as well as*

- *$\alpha_1 = 0.5$, $\beta_1 = 0.7$, $c_{\mathcal{A}}(a_1) = 0$, i.e. action $a_1$ is free with small drift towards $0$,*
- *$\alpha_2 = 0.5$, $\beta_2 = 0.9$, $c_{\mathcal{A}}(a_2) = 10$, i.e. action $a_2$ is costly with higher drift towards $0$.*

*Both the value iteration and the policy iteration algorithm deliver the optimal policy $u^*$ given by*

$$u^*(x) = \begin{cases} a_1 & \text{for } x \leq 15, \\ a_2 & \text{for } x > 15, \end{cases}$$

*and the value function illustrated in Figure 1.3. On average, the policy iteration is approximately 100 times faster than the value iteration.*

### Relation to other optimization methods

Why is value iteration resp. policy iteration the right ansatz? Is it possible to find the optimal policy by other common optimization algorithms such as linear programming or the gradient-descent method? These questions will be answered in the following.

The problem of computing the optimal policy for a MDP with the discounted-cost criterion can indeed be formulated as a linear programming problem [14]. Then the simplex method can be used to calculate the optimal control policy [79]. However, linear programming does not make use of the special structure that is given by a Markov control problem [43] such that policy iteration is regarded as the more efficient approach [31].

[2] As the state $x = 0$ is absorbing, the long-term average costs are given by $c_{\mathcal{S}}(0) = 0$ independent of the chosen action. This is why it only makes sense to consider discounted costs.

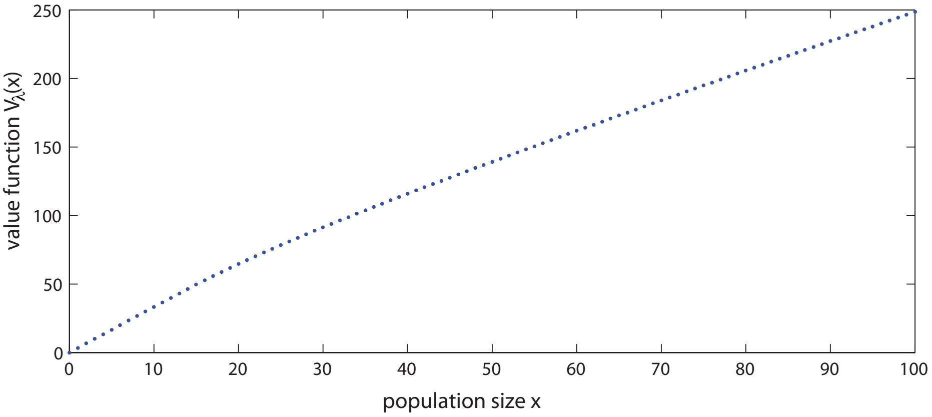

Figure 1.3: **Value function $V_\lambda(x)$.** Function of optimal discounted costs for the controlled population process of Example 1.18. The respective cost function is given by $c(x,a) = c_\mathcal{S}(x) + c_\mathcal{A}(a)$ with $c_\mathcal{S}(x) = x$.

As for the application of the gradient-descent method we make the following consideration. The given optimization problem consists of finding a policy minimizing the cost functional $J_\lambda$ resp. $\bar{J}$. The impact of the policy onto the cost functional, however, is quite complex and somehow indirect: The policy determines the generator and the cost function which enter the cost functional. In general, there might be no systematic structure in this relationship. Instead, the impact on the generator and the impact on the cost function can be completely independent of each other. For the application of the gradient-descend method one would have to assume that, for each state $x \in \mathcal{S}$, the action $a = u(x)$ defines a continuous parameter and that the cost functional is differentiable with respect to this parameter.
For an illustration, we give an idea of how a gradient-descent could look like for the 2-state-example 1.2. Given the two actions $a_1, a_2$, we consider some kind of linear interpolation: For $0.01 \leq a \leq 0.1$ define

$$L_a = \begin{pmatrix} -a & a \\ a & -a \end{pmatrix}$$

and $c(x,a) = c_\mathcal{S}(x) + \frac{2}{0.1-0.01} \cdot (a - 0.01)$ such that for $a = 0.01$ resp. $a = 0.1$ we get exactly the generator and the cost function defined in Example 1.2. Calculating the cost functional $J_\lambda$ and its gradient for state $x_2$ and all policies delivers Figure 1.4. Here, the gradient-descent method would deliver the optimal policy $u^*(x_1) = 0.01$, $u^*(x_2) = 0.1$.

In some sense, the policy iteration approach can be seen as a discrete, pointwise analogue to the gradient-descent method: Given a policy, the algorithm looks pointwise (i.e. for each state separately) for a maximal improvement.

We can conclude that, from all mentioned numerical approaches it is the policy iteration method which is best suited to solve a Markov control problem. In practice, policy iteration is widely used and shows a satisfying performance [79].

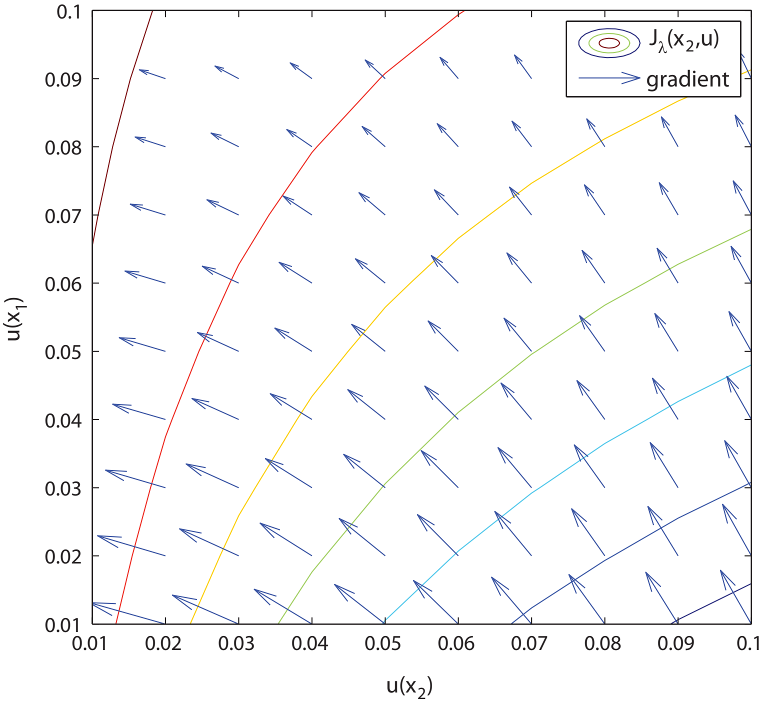

Figure 1.4: **Gradient descent.** Cost functional $J_\lambda$ and its gradient for state $x_2$ and all policies of the 2-state-example 1.2.

So far, we considered Markov decision processes that are completely observable at all times and permit an instantaneous control adaption. Situations of limited state information are analyzed in the theory of partially observable Markov decision processes which will be presented in the following section.

## 1.3 Partially Observable Markov Decision Processes

A partially observable Markov decision process (POMDP) can be seen as a generalization of a Markov decision process allowing uncertainty with respect to the state of the process. While a (standard) Markov decision process is assumed to be completely observable at all times, the information about the state of a POMDP may be incomplete. There exist many application areas for POMDPs such as machine maintenance or quality control.

A lot of research has been done on POMDPs, see for example the survey paper by Monahan [46]. In the following, we will give a short overview about the theory of POMDPs for a discrete state space and discrete time, considering the discounted-cost criterion. In general, the ideas can be transferred to more general state spaces [49]. In the end of this section, we describe further approaches for Markov control with incomplete information, particularly in the context of machine maintenance.

We consider the discrete-time Markov control model

$$\big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{P_a : a \in \mathcal{A}\},\, \tilde{c}\,\big),$$

see Remark 1.3, where $\mathcal{A}(x) = \mathcal{A}$ for all $x \in \mathcal{S}$. The corresponding Markov decision process $(X_n)_{n\in\mathbb{N}_0}$ will here be called the *core process*. We assume both $\mathcal{S}$ and $\mathcal{A}$ to be finite.
In the theory of POMDPs the states of the core process cannot be observed directly. Instead, there is another stochastic process $(O_n)_{n\in\mathbb{N}_0}$ associated with $(X_n)_{n\in\mathbb{N}_0}$ which takes on values in a denumerable set $\mathcal{O}$ called the *observation space*. The *observation process* $(O_n)_{n\in\mathbb{N}_0}$ delivers an indirect information about the core process. More precisely, the relation between $(X_n)_{n\in\mathbb{N}_0}$ and $(O_n)_{n\in\mathbb{N}_0}$ is defined by an *information rule* $q : \mathcal{S} \times \mathcal{O} \to [0, 1]$ with

$$q(x, o) = \mathbb{P}(O_n = o|X_n = x),$$

i.e. given that the state of the core process is $X_n = x \in \mathcal{S}$, an observation $o \in \mathcal{O}$ is induced with probability $q(x, o)$. Just like the transition matrix $P_a$ and the cost function $\tilde{c}$, also the information rule may depend on the action. In this case we write $q_a(x, o)$ for $a \in \mathcal{A}$. Moreover, given $a \in \mathcal{A}$, let $Q_a = (q_a(x, o))_{x\in\mathcal{S},o\in\mathcal{O}}$ be the information matrix.
Without loss of generality, we assume that the costs produced by the process are not observable during the control procedure. Observable costs would give an additional information about the process and have to be considered as a part of the observation process [63].

The POMDP can be converted into an equivalent completely observable MDP by considering the set

$$\Psi(\mathcal{S}) := \{\psi \in \mathbb{R}^{|\mathcal{S}|} : \sum_{x\in\mathcal{S}} \psi(x) = 1,\, \psi(x) \geq 0\ \forall x \in \mathcal{S}\}$$

of probability measures on $\mathcal{S}$ as a new state space [55]. To this end, let $H_n := (\psi_0, A_0, O_1, A_1, ..., A_{n-1}, O_n)$ be the series of observations and chosen actions up to time $n \in \mathbb{N}$, where $\psi_0$ is a given initial distribution on $\mathcal{S}$. $H_n$ represents the information available at time $n \in \mathbb{N}$. For each $n \in \mathbb{N}$, the history $H_n$ depends on the core process $(X_k)_{k=0,...,n}$ up to this time and is therefore a random variable itself. We define the process $(\psi_n)_{n\in\mathbb{N}}$ of so called *information vectors* on the new state space $\Psi(\mathcal{S})$ by

$$\psi_n(x) := \mathbb{P}(X_n = x|H_n),$$

where generally $\mathbb{P}(B|Y) := \mathbb{E}(\mathbb{1}_B|\sigma(Y))$ for a measurable set $B$ (element of a given $\sigma$-algebra) and a random variable $Y$. Given $\psi_n$ as well as $A_n = a$ and $O_{n+1} = o$, the next information vector $\psi_{n+1}$ is determined by Bayes' formula and the law of total probability:

$$\begin{aligned}\psi_{n+1}(x) &= \mathbb{P}(X_{n+1} = x|\psi_n, A_n = a, O_{n+1} = o) \\ &= \frac{\mathbb{P}(O_{n+1} = o|X_{n+1} = x, \psi_n, A_n = a) \cdot \mathbb{P}(X_{n+1} = x|\psi_n, A_n = a)}{\mathbb{P}(O_{n+1} = o|\psi_n, A_n = a)} \\ &= \frac{q_a(x, o) \cdot \sum_{y \in \mathcal{S}} P_a(y, x)\psi_n(y)}{\sum_{\tilde{y} \in \mathcal{S}} q_a(\tilde{y}, o) \cdot \sum_{y \in \mathcal{S}} P_a(y, \tilde{y})\psi_n(y)} =: T_x(\psi_n|a, o). \qquad (1.19)\end{aligned}$$

For notational convenience, we define $T(\psi|a, o) := (T_x(\psi_n|a, o))_{x \in \mathcal{S}}$ (such that $\psi_{n+1} = T(\psi_n|a, o)$) and

$$\gamma(o|\psi, a) := \sum_{\tilde{y} \in \mathcal{S}} q_a(\tilde{y}, o) \cdot \sum_{y \in \mathcal{S}} P_a(y, \tilde{y})\psi(y)$$

which refers to the denominator in equation (1.19).

The vector $\psi_n$ summarizes all the information available at time $n \in \mathbb{N}$, see e.g. [7, 63]. As $\psi_n$ is a function of the (random) history $H_n$, the series $(\psi_n)_{n \in \mathbb{N}_0}$ is again a stochastic process. For each sequence $a_0, a_1, ... \in \mathcal{A}$ of actions, this process fulfills the Markov property (see e.g. [30]), i.e. it holds

$$\mathbb{P}(\psi_{n+1} \in B|\psi_0, ..., \psi_n, a_n) = \mathbb{P}(\psi_{n+1} \in B|\psi_n, a_n) \quad \forall n \in \mathbb{N}_0, B \in \mathcal{B}(\Psi(\mathcal{S})),$$

where $\mathcal{B}(\Psi(\mathcal{S}))$ is the Borel-$\sigma$-algebra on $\Psi(\mathcal{S})$. Note that the state space $\Psi(\mathcal{S})$ of this Markov process is not denumerable anymore.

The cost functional associated with the process $(\psi_n)$ is defined by

$$C(\psi, a) := \sum_{x \in \mathcal{S}} \psi(x)\tilde{c}(x, a) = \mathbb{E}\big(c(X_n, a)|\psi\big),$$

denoting the expected costs given the information $\psi$. A (deterministic stationary) policy is a function

$$u : \Psi(\mathcal{S}) \to \mathcal{A},$$

where $u(\psi)$ denotes the action to chose when the current information is $\psi$. The set of all policies is denoted by $\mathcal{U}$. Given a policy $u \in \mathcal{U}$, the system evolves according to the following procedure, see Figure 1.5 for a schematic representation. Given the information $\psi_n$ at time $n \in \mathbb{N}$ (corresponding to a hidden state $X_n = x \in \mathcal{S}$ and an observation $O_n = o \in \mathcal{O}$), an action $A_n = a \in \mathcal{A}$ is chosen according to $u$. We have to pay the costs $C(\psi, a)$, and the next state $X_{n+1}$ of the core process is determined according to the transition rule $P_a$. Instead of $X_{n+1}$ we get an observation $O_{n+1}$ which is generated by $Q_a$. This observation together with $\psi_n$ and $a$ determines the next information $\psi_{n+1}$, and the procedure restarts.

Given a policy $u \in \mathcal{U}$ and an initial information $\psi \in \Psi(\mathcal{S})$, consider the discounted-cost criterion defined by

$$J_\lambda(\psi, u) = \mathbb{E}^u_\psi \left( \sum_{n=0}^{\infty} e^{-\lambda n} C(\psi_n, u(\psi_n)) \right), \qquad (1.20)$$

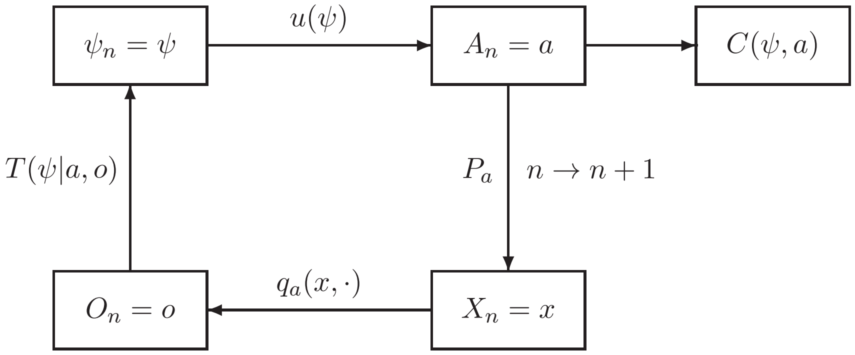

Figure 1.5: **Schematic representation of a POMDP.**

where $\lambda > 0$ is a discount factor as in Section 1.1. The optimal value function, defined by

$$V_\lambda(\psi) := \inf_{u \in \mathcal{U}} J_\lambda(\psi, u), \quad \psi \in \Psi(\mathcal{S}),$$

fulfills the Bellman equation

$$V_\lambda(\psi) = \inf_{a \in \mathcal{A}} \left\{ C(\psi, a) + e^{-\lambda} \sum_{o \in \mathcal{O}} V_\lambda(T(\psi|a, o))\gamma(o|\psi, a) \right\}, \tag{1.21}$$

see [46]. For a finite action space there exists an optimal policy $u^*$ [55] which minimizes the right hand side of (1.21). Other criteria which guarantee the existence of an optimal policy are given in [30].

A typical example for a POMDP is the following.

**Example 1.19** (Quality control/Machine Maintenance)**.** *Consider a machine which can be either in a good condition (state $x_1$) or in a bad condition (state $x_2$). The state of the machine cannot be observed directly. Instead we get an indirect information by observing the machine's output which can have a good quality (observation $o_1$) or a bad quality (observation $o_2$). There are three possible actions available: $a_1$ stands for doing nothing (produce without inspection), $a_2$ stands for inspecting the machine, and $a_3$ stands for repairing the machine. Let the corresponding transition matrices and information matrices be given by*

$$P_1 = \begin{pmatrix} 1-\gamma & \gamma \\ 0 & 1 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_2 & \alpha_2 \end{pmatrix},$$

$$P_2 = \begin{pmatrix} 1-\gamma & \gamma \\ 0 & 1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} \beta_1 & 1-\beta_1 \\ 1-\beta_2 & \beta_2 \end{pmatrix},$$

$$P_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad Q_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

*where* $0 \leq \gamma, \alpha_1, \alpha_2, \beta_1, \beta_2 \leq 1$. *The probability of the machine to deteriorate is given by* $\gamma$, *while* $\alpha_1$, $\alpha_2$ *resp.* $\beta_1$, $\beta_2$ *describe the grade of information when doing nothing resp. when inspecting the machine: For example,* $\alpha_1 = 1$, $\alpha_2 = 0$ *and* $\beta_1 = \beta_2 = 1$ *means that there is no information for action* $a_1$ *(doing nothing) and full information for action* $a_2$ *(inspecting the machine). The optimal policy with respect to these parameters and the discounted-cost criterion* (1.20) *is given in Figure 1.6. The choice of* $P_3$ *implies that after repairing the machine it will almost surely be in a good condition and* $Q_3$ *represents complete information.*

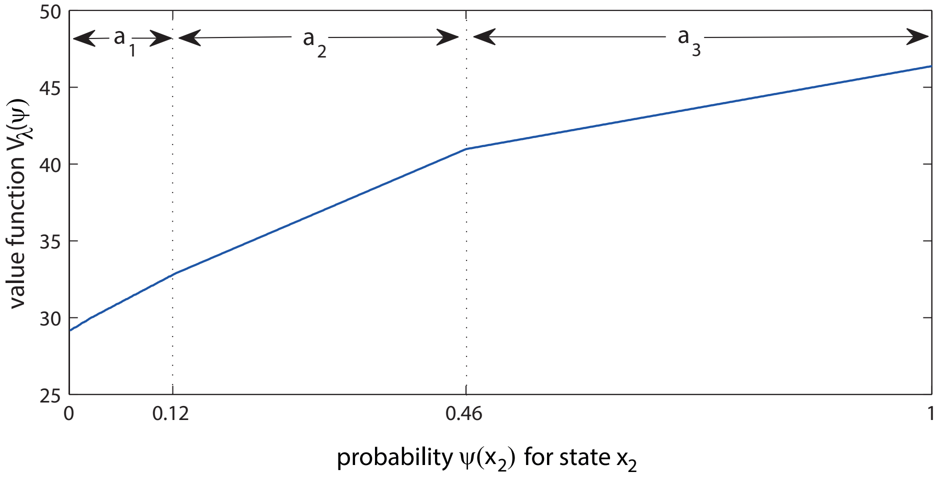

Figure 1.6: **Value function and optimal policy for machine maintenance.** Function $V_\lambda(\psi)$ of optimal discounted costs and related optimal policy for Example 1.19 given that $\lambda = 0.1$, $\gamma = 0.1$, $\alpha_1 = 1$, $\alpha_2 = 0$, $\beta_1 = \beta_2 = 1$ and $c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$ with $c_{\mathcal{S}}(x_1) = 0$, $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_1) = 0$, $c_{\mathcal{A}}(a_2) = 2$, $c_{\mathcal{A}}(a_3) = 10$. For $\psi(x_2) < 0.12$, which refers to a low probability for the machine to be in a bad condition, the optimal action to choose is $a_1$ (do nothing); for $0.12 \leq \psi(x_2) < 0.46$ the optimal choice is given by $a_2$ (inspect the machine); and for $\psi(x_2) \geq 0.46$ the optimal choice is given by $a_3$ (repair the machine).

### Numerics

We have seen that a POMDP with state space $\mathcal{S}$ can be turned into a completely observable MDP on the information space $\Psi(\mathcal{S})$. This suggests to apply the algorithms described in Section 1.2 in order to calculate an optimal policy for a POMDP. The arising difficulty is that the new state space is continuous. One way to handle this new setting is to simply discretize the set $\Psi(\mathcal{S})$. However, it is not clear how the discretization should be chosen and under which conditions it delivers a good approximation of the exact value function resp. optimal policy.

It is still possible to use dynamic programming for POMDPs. This is due to the fact that the value function can be approximated arbitrarily closely by piecewise linear and concave functions and that the value iteration step (see Section 1.2) preserves these properties [28, 62].[3] See Figure 1.6 for an illustration. The concavity of the value function (as a function of the information $\psi$) has an intuitive interpretation: In the "middle" of the information space $\Psi(\mathcal{S})$ we are concerned with a high uncertainty about the true state of the system which prevents from choosing the action appropriately and results in higher costs. The outer boundaries of $\Psi(\mathcal{S})$, however, refer to high exactness and allow for adequate control.
A piecewise linear and concave function $f$ can be represented by a finite set $\mathcal{V}$ of $|\mathcal{S}|$-dimensional vectors by writing

$$f(\psi) = \min_{v \in \mathcal{V}} \langle \psi, v \rangle,$$

where $\langle \cdot, \cdot \rangle$ again refers to the standard inner product on $\mathbb{R}^{|\mathcal{S}|}$. Each vector $v \in \mathcal{V}$ corresponds to the choice of an action, and therefore each step in the value iteration approach consists of finding a suitable set $\mathcal{V}$. Different algorithms have been developed to perform this step [9, 10, 46]. The set $\mathcal{V}^*$ corresponding to the optimal policy possibly contains an exponential number of vectors which implies an exponential runtime of the related algorithms. Given the set $\mathcal{V}^*$ (or at least an approximation), the final step in this approach is to extract the corresponding policy. How this can be done is described in [28, 33] and others.
Beside the value iteration approach there also exists an ansatz for finding the optimal policy by search in policy space. Hansen [28] proposes an algorithm which clearly outperforms the value iteration in several examples.

In summary one can state that problems of POMDP are more complex to solve than problems assuming full observability which is not surprising due to the more complex structure of the control model.

### Machine maintenance as a special setting

Different variants of the quality control example 1.19 as a special case of a POMDP are discussed by several authors. Ross [51] considers the situation given in Example 1.19 with $\alpha_1 = 1$, $\alpha_2 = 0$ (no information without inspection) and $\beta_1 = \beta_2 = 1$ (full information with inspection) and discovers that the optimal policy is always of the form illustrated in Figure 1.7. Note that for the case of imperfect information (i.e. $\beta_1, \beta_2 \neq 1$) the optimal policy can have a much more complicated structure [46].

Rosenfield [50] generalizes the two-state results presented by Ross to the case of an arbitrary finite number of states. As a new state space Rosenfield considers the set of pairs $(x, k)$, where $x \in \mathcal{S}$ is the last state of the process known exactly to the controller and $k \in \mathbb{N}_0$ is the number of time steps since this state was observed.

[3] The authors consider the problem of maximizing rewards instead of minimizing costs which implies convexity instead of concavity. We translated their results to our situation of minimizing costs.

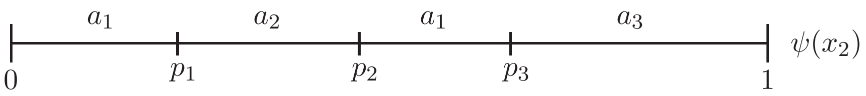

Figure 1.7: **Structure of the optimal policy for machine maintenance.** Given Example 1.19 with $\alpha_1 = 1$, $\alpha_2 = 0$, $\beta_1 = \beta_2 = 1$, the optimal control policy always has the indicated structure [51]. If the probability $\psi(x_2)$ is smaller than $p_1$ or in between $p_2$ and $p_3$, the optimal choice of action is $a_1$ (produce without inspection); for $p_1 < \psi(x_2) < p_2$ it is recommended to choose $a_2$ (make an inspection); for $\psi(x_2) > p_3$, the optimal choice is given by $a_3$ (repair the machine).

A structure of the optimal policy equivalent to the one of the two-state-example is discovered. In [72], White gives similar results but for weaker conditions. All of them consider both the discounted- and the average-cost criterion.

All contributions to this problem of machine maintenance have in common that the state space exhibits an ordered structure (measuring in some sense the degree of deterioration of the considered object) and that the natural dynamics are nondecreasing in the sense that the system cannot improve its status without any intervention by the controller. Especially the last state (failure) is assumed to be absorbing. Possible interaction consists of sending the system back to the initial state (repairing).

### Continuous-time Markov decision processes with limited state information

So far the mentioned literature referred to discrete-time POMDPs. Models with a continuous time parameter were examined (among others) by the following authors.

Bather [4] describes an optimal stopping problem for a Brownian motion which cannot be observed directly. Instead, the decision maker may "buy" some incomplete information about the process. The problem is converted into a completely observable control process on the set of normal distributions which is then divided into three regions corresponding to the choice of possible actions: (i) do not interact, (ii) collect information, (iii) stop the process. The goal is to declare the regions such that the expected costs (which consist of the total information costs and the final costs when stopping the process) are minimized. This is done by formulating partial differential equations which describe the boundaries of the optimal regions. The author does not give any reference to the theory of POMDPs.

A similar model is given by Anderson and Friedman in [1]. They extend the theory to situations where – instead of stopping the process – one can bring it back to the initial state, and consider the case of defective observations.

In [54] Savage studies surveillance problems for production processes. The considered process is assumed to be a Poisson process (implying continuous time, discrete state space) which can be observed by making a costly inspection. After such an inspection (which is assumed to give perfect information about the state of the process) it has to be decided whether to let the process continue or to interact by making a repair which sends the process back to some initial state. In this setting a policy consists of declaring a continuation region as well as a state dependent time for the next observation. A recursive equation for the optimal income is given, but without reference to dynamic programming resp. the Bellman equation.

More recent work has been carried out by Kim [34]. The author considers a continuous-time process of deterioration similar to the discrete settings of machine maintenance described above. The system is assumed to be nondecreasing in the set of states "healthy", "warning" and "failure" and can be influenced by choosing one of three actions (do nothing, inspect or interact). A policy is defined as a function of the probability of being in the warning state, and the optimal policy is characterized by critical thresholds for this probability. The state of the process is determined by taking a sample which is only possible at predefined equidistant discrete points in time. Other recent works considering the inspection times as a part of a maintenance problem are given by [17, 70] and others.

In [80] a medical context is examined. Here the process of disease detection is described by considering the three states "healthy", "having undiagnosed disease" and "having diagnosed disease". The goal is to find an optimal arrangement of a fixed number of examinations within a fixed time interval.

As in the discrete-time setting all these approaches assume a special structure in the state space and only allow for a restricted type of interaction. The aim of this thesis is to find a more general setting which avoids the interpretation of deterioration and repair/replacement in a maintenance problem. To this end, we will in the following Chapter 2 develop a modified control model for continuous-time Markov decision processes that can be observed at discrete but flexible points in time. The state and action space will be free of interpretation, and the dynamics will have no special structure.

Chapter 2

# Markov Control with Information Costs

A Markov decision process is a stochastic process that can be controlled by manipulating its transition rates within a given framework. The process and its control produce costs according to a given cost function, and the goal is to find a control rule which minimizes a fixed cost criterion. How such an optimization problem can be formulated and solved has been described in Chapter 1. We also discussed situations of limited state information where the process is not completely observable. The different approaches to handle such a situation lead to a significant extension of the state space (compare the theory of POMDPs) or assume the state and action space to have a special ordered structure.

In this chapter we will introduce a novel model for continuous-time Markov decision processes on a discrete state space which are not permanently observable. Instead, each observation of the process produces costs which enter the considered cost functional. The resulting control problem consists of finding for each state not only an adequate action but also a date for the next examination of its state. The situation resembles the machine maintenance problems described in Section 1.3, however we will formulate the problem for a general state and action space without fixing any special characteristics of the considered dynamics. This general setting enables both a thorough theoretical analysis and an interesting real-world application in the medical context (see Chapter 3).

We will first establish in Section 2.1 the modified Markov control model and define the corresponding control procedure. Of central relevance will be the parameter $k_{\text{info}}$ of so called *information costs* which arise each time the state of the process is detected. We will reformulate the two optimality criteria of discounted and average costs for the new setting and give a detailed analysis for both of them in Sections 2.2 and 2.3. This analysis will contain the derivation of a modified Bellman equation and the calculation of a cost splitting where the total costs are divided into components of state costs, action costs and information costs. Furthermore, we will study how the optimal policy and the optimal costs depend on the information cost parameter $k_{\text{info}}$ and how deviations from the optimal examination times influence the overall costs.

## 2.1 The Control Model

In line with the notation of section 1.1, we define the *Markov control model with information costs* to be a tuple

$$\Big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{L_a : a \in \mathcal{A}\},\, c\,,\, k_{\text{info}}\Big). \tag{2.1}$$

As before, $\mathcal{S}$ is the set of states which is assumed to be denumerable. $\mathcal{A}$ is a Borel space of actions and $\mathcal{A}(x)$ denote the set of actions available in state $x \in \mathcal{S}$. For simplicity, we assume $\mathcal{A}(x) = \mathcal{A}$ for all $x \in \mathcal{S}$ which is no decisive restriction. Moreover, $L_a$ is the generator describing the dynamics of the process given action $a \in \mathcal{A}$, fulfilling $L_a(x, y) \geq 0$ for all $x \neq y$ as well as $L_a(x, x) = -\sum_{y \neq x} L_a(x, y)$. We set $l_a(x) := -L_a(x, x)$ and assume $\sup_{a\in\mathcal{A}} l_a(x) < \infty$ for all $x \in \mathcal{S}$. The function $c : \mathcal{S} \times \mathcal{A} \to [0, \infty)$ defines the costs produced by the process per unit of time depending on the state and the chosen action. The new parameter

$$k_{\text{info}} > 0$$

is a constant number denoting the price for a state observation. That is, each time we determine the state of the system we have to pay the fee $k_{\text{info}}$.

More precisely, the control procedure is the following. Starting with some (known) state $X_{t_0} = x_0 \in \mathcal{S}$ at time $t_0 = 0$, one has to choose not only an action $a \in \mathcal{A}(x_0)$ but also a time $t_1 = t_0 + \tau > t_0$ for the next state observation. Within the time interval $(t_0, t_1)$ the process $(X_t)_{t\geq 0}$ evolves according to the generator $L_a$ and produces costs according to $c(\cdot, a)$. This evolution and the arising costs cannot be observed, we only determine the state $X_{t_1}$ at time $t_1$ by making a test which produces costs $k_{\text{info}}$. Given the state of the process at time $t_1$, the procedure restarts.
The resulting *observation times* $(t_j)_{j\in\mathbb{N}_0}$ which are recursively determined by this procedure identify the moments in time where the state of the process is observed and a decision has to be taken. We call the related process of observations $(X_{t_j})_{j\in\mathbb{N}_0}$ the *observation process.*

The following assumption basically determines the setting and will permit a clear and comprehensible formulation of the control problem.

**Assumption 2.1.**
a) The test is assumed to give instantaneous and full information such that the state of the process at the observation times is known with certainty.
b) During a time interval $(t_j, t_{j+1})$ the action is constant, i.e. it cannot be changed blindly without making a test to determine the state.

We now define the term "policy" within the new setting. Knowing from the original Markov control theory that a restriction to deterministic, time homogeneous policies has no further significance, we directly focus on this class of control rules.

**Definition 2.2** (Markov policy)**.** Given the Markov control model with information costs (2.1), a deterministic, time homogeneous Markov policy is a function

$$u : \mathcal{S} \to \mathcal{A} \times (0, \infty], \quad u(x) = (a(x), \tau(x)),$$

declaring for each state $x \in \mathcal{S}$ an action $a(x) \in \mathcal{A}$ as well as a *lag time* $\tau(x) \in (0, \infty]$ defining the time length for the next period of hidden progress.
We denote the set of all these policies by $\mathcal{U}$.

We explicitly allow the parameter $\tau(x)$ to be **infinite**. The choice of $\tau(x) = \infty$ has a reasonable interpretation: It simply means to make no further tests at all, but to let the process run under constant control forever. We set

$$e^{-\lambda\infty} := 0 \quad \text{and} \quad e^{L_a\infty} := \lim_{t\to\infty} e^{L_a t}, \tag{2.2}$$

assuming that this limit exists; otherwise we set $e^{L_a\infty} := I$.[4] By this definition, all analytic expressions will have a straightforward interpretation for cases of infinite lag times.
In contrast, the value $\tau(x) = 0$ is excluded by the following argument: A vanishing lag time would mean to immediately repeat a state test, which delivers no further information but produces additional costs $k_{\text{info}} > 0$ and is therefore not reasonable.

Given an initial distribution $\nu$ on $\mathcal{S}$, a policy $u$ defines a probability measure $\mathbb{P}^u_\nu$ on the set of possible state-action-realizations

$$\big\{(X_t, A_t)_{t\geq 0} : \ X_t \in \mathcal{S}, A_t \in \mathcal{A} \ \forall t \geq 0\big\}$$

by

- $\mathbb{P}^u_\nu(X_0 = x) := \nu(x)$,
- $t_0 := 0$ and $t_{j+1} := t_j + \tau(X_{t_j})$ for $j \in \mathbb{N}_0$,
- $A_t := a(X_{t_j})$ for all $t_j \leq t < t_{j+1}$, $t_j < \infty$,
- $\frac{\partial}{\partial t}\mathbb{P}^u_\nu(X_t = x | A_t = a) = \sum_{y\in\mathcal{S}} L_a(y, x)\mathbb{P}^u_\nu(X_t = x | A_t = a)$,

where $x \in \mathcal{S}$, $a \in \mathcal{A}$. Note that the observation times $t_j$ are random variables which may have the value $\infty$; in this regard we interpret $t_j + \infty := \infty$ for $t_j < \infty$ as well as $\infty + \infty := \infty$.
Again, we write $\mathbb{P}^u_x$ for $\nu = \delta_x$ and denote the corresponding expectation values by $\mathbb{E}^u_\nu$ resp. $\mathbb{E}^u_x$.

Figure 2.1 illustrates the controlled process $(X_t)_{t\geq 0}$ for a fixed policy $u \in \mathcal{U}$.

[4] In fact, the definition of $e^{L_a\infty}$ is of no further significance as this expression will always be multiplied by $e^{-\lambda\infty} = 0$.

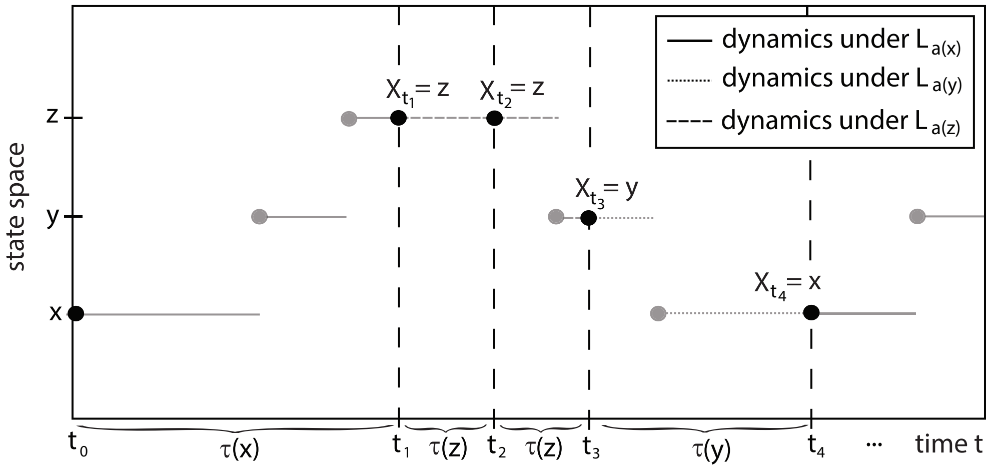

Figure 2.1: **Controlled Markov process with information costs.** Possible trajectory of a Markov decision process with information costs given a deterministic stationary policy $u$. Starting in a (known) state $x \in \mathcal{S}$ at time $t_0$, the dynamics of the process during the time interval $[t_0, t_0+\tau(x))$ are determined by the generator $L_{a(x)}$, i.e. the process stays in $x$ for some random period of time which is exponentially distributed with parameter $l_{a(x)}(x) = -L_{a(x)}(x,x)$ and then jumps to a state $y \neq x$ with probability $\frac{L_{a(x)}(x,y)}{l_{a(x)}(x)}$. However, these dynamics are unobserved which is illustrated by the transparency of the corresponding lines and dots. We only get a pointwise information about the state of the process at time $t_1 = t_0 + \tau(x)$. Given this state $X_{t_1} = z$, the control is adapted and the procedure restarts.

### Optimality criteria

We are interested in formulating optimality criteria which measure not only the process costs defined by the cost function $c$ but also the costs for making the tests. While the process costs arise continuously over time and depend on the evolution of the process, the information costs $k_{\text{info}}$ are a fixed constant which has to be paid instantaneously when making a test. As in Section 1.1 we consider discounted costs and average costs for an infinite time horizon.

a) The *expected discounted-cost criterion*: Given an initial state $x \in \mathcal{S}$ and a discount factor $\lambda > 0$, the total expected discounted costs under control $u \in \mathcal{U}$ are defined by

$$J_\lambda(x,u) := \mathbb{E}_x^u \left( \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<\infty}} \left( \int_{t_j}^{t_{j+1}} e^{-\lambda s} c(X_s, a(X_{t_j}))\, ds + e^{-\lambda t_{j+1}} k_{\text{info}} \right) \right). \quad (2.3)$$

The corresponding *optimal discounted-cost function* or *value function of discounted costs* is given by

$$V_\lambda(x) := \inf_{u\in\mathcal{U}} J_\lambda(x,u). \quad (2.4)$$

b) The *expected average-cost criterion*: Given an initial state $x \in \mathcal{S}$, the long-run expected average costs under control $u \in \mathcal{U}$ are defined by

$$\bar{J}(x,u) := \limsup_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<T}} \left( \int_{t_j}^{t_{j+1}\wedge T} c(X_s, a(X_{t_j}))\,ds + k_{\text{info}} \right) \right), \quad (2.5)$$

where $t_{j+1} \wedge T := \min\{t_{j+1}, T\}$. The corresponding *optimal average-cost function* or *value function of average costs* is given by

$$\bar{V}(x) := \inf_{u\in\mathcal{U}} \bar{J}(x,u). \quad (2.6)$$

Note that in both definitions the function $c$ is evaluated in the first argument at $X_s$ with $s$ running over time, while in the second argument it is evaluated at $a(X_{t_j})$ with $t_j$ fixed for each interval. This follows from the fact that the state (which we do not observe during such an interval) changes as usual, while the action stays the same.

The definition of optimal (resp. $\varepsilon$-optimal) policies is analogous to the one given in the original setting, see page 10.

**Remark 2.3.** *In the case of finite lag times $\tau(x)$ (resulting in finite testing times $(t_j)_{j=0,1,\ldots}$) we can rewrite the average-cost criterion* (2.5) *as*

$$\bar{J}(x,u) = \limsup_{n\to\infty} \mathbb{E}_x^u \left( \frac{1}{t_n} \sum_{j=0}^{n-1} \left( \int_{t_j}^{t_{j+1}} c(X_s, a(X_{t_j}))\,ds + k_{\text{info}} \right) \right). \quad (2.7)$$

*For infinite $t_j$, however, this expression has no direct interpretation, which motivates to choose the more general notation given in equation* (2.5).
*Note further that the choice of the average-cost criterion* (2.5) *is not self-evident. Another idea would be to set, as for finite $t_j$,*

$$\tilde{J}(x,u) = \limsup_{n\to\infty} \mathbb{E}_x^u \left( \frac{1}{n} \sum_{j=0}^{n-1} \frac{1}{t_{j+1}-t_j} \left( \int_{t_j}^{t_{j+1}} c(X_s, a(X_{t_j}))\,ds + k_{\text{info}} \right) \right)$$

*which means to first calculate the average costs for each time interval – resulting in a new cost function depending on state and control – and afterwards calculate the average over all intervals. It turns out that the formulation of the Bellman equation in this case is much easier than for the function defined in* (2.5)*: We just have to consider a discrete-time process jumping between the states according to the transition probabilities of $P_{a,\tau} := \exp(L_a\tau)$ with a cost function defined by the average costs of the state. The problem is that in this scheme the total time spent in one state gets lost – and due to this the proportions of the total times for different states. That is, we only optimize the average costs and the transition probabilities, but not the relation of total times. For a "costly" state this can result in long residence times $\tau$ although it would be (judging according to our intuition) better to quickly switch to a better state.*

### Comparison to other control models

Before we start the analysis of the two cost criteria, we briefly explain how the presented model differs from other models of stochastic control. Regarding the **original Markov control theory**, the main difference is given by the observability of the process and the permanence of interaction: While a standard Markov decision process is permanently observed and controlled (see e.g. [5, 26, 48]), the evolution of the process in the new setting is most of the time hidden to the controller. Only at single discrete points in time the state is determined and the action is adapted. This is generated by the additional parameter of information costs $k_{\text{info}}$.

This difference in the time scales of the process itself (which is continuous in time) and its observation and control (which are discrete in time) also separates the new model from the theory of **partially observable Markov decision processes**, see e.g. [33, 46, 63]. As described in Section 1.3, the grade of information about a POMDP is defined by the chosen action, and the choice of actions takes place on the same (usually discrete) timescale as the underlying process.

In contrast to the setting of **machine maintenance** which deals with deteriorating systems, our control model (2.1) is not linked to any specific interpretation. The approaches in the theory of machine maintenance always make use of the special structure in the state space and dynamics (see e.g. [38, 51, 72]), such that they cannot be transferred to our general setting.

Another type of control problem is the so called **impulse control**, see e.g. [19, 37, 45]. Congruently with the control procedure in our setting, an impulse control takes place at single points in time – called the intervention times – while the underlying dynamics are continuous in time. Choosing these intervention times is part of the control problem and each interaction produces a fixed amount of “intervention costs”, just as in our model. However, the intervention times are allowed to be stopping times with respect to the underlying dynamics – which implies full information about the process evolution. In other words, the controlled process is assumed to be completely observable at all times, and the decision whether to interact or not at some point in time may depend on the state of the process at this time. Moreover, in a problem of impulse control, the interaction at the intervention times usually consists of shifting the process to another state. This is motivated by the typical application in portfolio management where the value of the portfolio is controlled by selling or buying assets. In our setting, however, the interaction is given by the choice of an action which determines the future dynamics of the process.

## 2.2 The Discounted-Cost Criterion

Given the Markov control model (2.1) with information cost parameter $k_{\text{info}}$, our aim is now to analyze the discounted-cost criterion

$$J_\lambda(x,u) = \mathbb{E}_x^u \left( \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<\infty}} \left( \int_{t_j}^{t_{j+1}} e^{-\lambda s} c(X_s, a(X_{t_j}))\, ds + e^{-\lambda t_{j+1}} k_{\text{info}} \right) \right)$$

defined in (2.3). The analysis will be restricted to **finite $\mathcal{S}$**.

In Section 2.2.1 we will reproduce the results of the original setting presented in Section 1.1.1: We will show that the new cost functional $J_\lambda(x,u)$ fulfills a recursion which can be interpreted as a fix-point equation of a suitable operator. This operator again satisfies special monotonicity properties and can be used to prove the discounted-cost optimality equation for the new setting. As before, this equation characterizes not only the value function of minimal costs but also the optimal policy. We will discuss the existence of an optimal policy as well as its determination by iterative methods.

The further analysis (Sections 2.2.2-2.2.4) has no analogue in the original setting; instead, the information cost parameter $k_{\text{info}}$ will play an essential role. First of all, we will calculate a "cost splitting" of the value function separating the optimal costs according to their origins. Especially, we are interested in the "net costs" – which are the total costs without the information costs caused by $k_{\text{info}}$ – as these are the basis for a comparison to the optimal costs of the original control problem.

Then we will analyze how the value function and the optimal policy depend on the parameter $k_{\text{info}}$: Is there some kind of monotonicity and/or continuity regarding the respective dependency?

Upon answering this question, we will close by a short sensitivity analysis for the lag time parameter $\tau$, checking how slight deviations from the optimal lag time influence the costs.

### 2.2.1 The Discounted-Cost Optimality Equation

Adapting the procedure of Section 1.1.1, we first investigate the cost functional $J_\lambda(x,u)$ for a given policy $u \in \mathcal{U}$, in order to then consider the value function of optimal costs and the related optimal policies.

#### Cost functional $J_\lambda(x,u)$ for a given policy $u \in \mathcal{U}$

For ease of notation we introduce a new cost function

$$C(x,a,\tau) := \mathbb{E}_x^a \left( \int_0^\tau e^{-\lambda s} c(X_s, a)\, ds \right)$$

denoting the expected costs during the time interval $(0,\tau)$, $\tau \in (0,\infty]$, of hidden progress when the process starts in $x \in \mathcal{S}$ and action $a \in \mathcal{A}$ is applied. By this definition we can write the cost functional given in (2.3) as

$$J_\lambda(x,u) = \mathbb{E}_x^u \left( \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<\infty}} e^{-\lambda t_j} \left( C(X_{t_j}, a(X_{t_j}), \tau(X_{t_j})) + e^{-\lambda\tau(X_{t_j})} k_{\text{info}} \right) \right). \tag{2.8}$$

This notation reveals the discrete nature of the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$. For a fixed policy $u \in \mathcal{U}$ we consider $J_\lambda(x,u)$ as a function of $x$ and use the notation

$$J_\lambda^u(x) := J_\lambda(x,u).$$

Similar to the results given in Section 1.1.1 we can decompose the sum in (2.8) to get a recursive equation for $J_\lambda^u$ and directly obtain

**Lemma 2.4.** *Given a policy $u \in \mathcal{U}$, the cost functional $J_\lambda^u$ fulfills the recursion*

$$J_\lambda^u(x) = C(x, a(x), \tau(x)) + e^{-\lambda\tau(x)} \left( k_{\text{info}} + (e^{L_{a(x)}\tau(x)} J_\lambda^u)(x) \right). \tag{2.9}$$

*Proof.* Straightforward decomposition of $J_\lambda^u$. □

The question is whether $J_\lambda^u$ is uniquely characterized by this recursion. As we fixed the state space $\mathcal{S}$ to be finite, the answer is yes. To see this, we analyze an operator $T_u : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ similar to the one considered in Section 1.1.1.

Given an action $a \in \mathcal{A}$ and a lag time $\tau \in (0,\infty]$, define the operator $T_{a,\tau} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ by

$$(T_{a,\tau}J)(x) := \begin{cases} C(x,a,\tau) + e^{-\lambda\tau}\Big(k_{\text{info}} + (e^{L_a\tau}J)(x)\Big), & \text{if } \tau < \infty, \\ C(x,a,\infty), & \text{if } \tau = \infty. \end{cases} \tag{2.10}$$

In line with this, we define the operator $T_u : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ for a policy $u(x) = (a(x), \tau(x))$ by

$$(T_uJ)(x) := \big(T_{u(x)}J\big)(x) = \big(T_{a(x),\tau(x)}J\big)(x). \tag{2.11}$$

Moreover, let the transition matrix $P_{a,\tau} \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ for an action $a \in \mathcal{A}$ and a lag time $\tau \in (0,\infty]$ be defined by

$$P_{a,\tau}(x,y) := \begin{cases} \left(e^{L_a\tau}\right)(x,y), & \text{if } \tau < \infty, \\ \lim_{t\to\infty} \left(e^{L_a\cdot t}\right)(x,y), & \text{if } \tau = \infty. \end{cases} \tag{2.12}$$

For a policy $u \in \mathcal{U}$, define the transition matrix $P_u$ by

$$P_u(x,y) := P_{u(x)}(x,y) = P_{a(x),\tau(x)}(x,y). \tag{2.13}$$

This transition matrix defines the dynamics of the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ given that the underlying process $(X_t)_{t\geq 0}$ is steered by the policy $u$.

By this notation, equation (2.9) can be seen as a fix-point equation, namely

$$J_\lambda^u = T_u J_\lambda^u.$$

Again, we can argue that the function $J_\lambda^u$ is uniquely characterized by this equation because the operator $T_u$ is contractive.

**Lemma 2.5.** *For every policy $u \in \mathcal{U}$, the operator $T_u$ defined in* (2.11) *is a contraction on* $(\mathbb{R}^{|\mathcal{S}|}, ||\cdot||_\infty)$.

*Proof.* For $J, \tilde{J} \in \mathbb{R}^{|\mathcal{S}|}$ and $\alpha := \max_{x\in\mathcal{S}} e^{-\lambda\tau(x)}$ (satisfying $0 \le \alpha < 1$) it holds that

$$\begin{aligned}
||T_u J - T_u \tilde{J}||_\infty &= \max_{x\in\mathcal{S}} \left|(T_u J)(x) - (T_u \tilde{J})(x)\right| \\
&= \max_{x\in\mathcal{S}} \left|\left(e^{-\lambda\tau(x)} e^{L_{a(x)}\tau(x)}(J - \tilde{J})\right)(x)\right| \\
&\le \alpha \cdot \max_{x\in\mathcal{S}} \left|\left(e^{L_{a(x)}\tau(x)}(J - \tilde{J})\right)(x)\right| \\
&= \alpha \cdot \max_{x\in\mathcal{S}} \left|\left(P_u(J - \tilde{J})\right)(x)\right| \\
&= \alpha \cdot ||P_u(J - \tilde{J})||_\infty \\
&\le \alpha \cdot ||J - \tilde{J}||_\infty,
\end{aligned}$$

where the last step follows from the fact that $P_u$ is a stochastic matrix. Note that, by $e^{-\lambda\infty} := 0$, this calculation is valid for the case $\tau(x) = \infty$ for one (or several) $x \in \mathcal{S}$, as well. □

In order to find a compact expression for $J_\lambda^u$ regarded as a vector in $\mathbb{R}^{|\mathcal{S}|}$, we introduce some notations that will be used again throughout this chapter.

Given a policy $u = (a(x), \tau(x))$, define the discount vector $e_\tau \in \mathbb{R}^{|\mathcal{S}|}$ by

$$e_\tau(x) := \begin{cases} e^{-\lambda\tau(x)}, & \text{if } \tau(x) < \infty, \\ 0, & \text{if } \tau(x) = \infty, \end{cases} \tag{2.14}$$

as well as a diagonal matrix $D_\tau \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ by

$$D_\tau(x,x) := e_\tau(x) \text{ and } D_\tau(x,y) := 0 \text{ for } x \neq y. \tag{2.15}$$

For the optimal policy $u^* = (a^*(x), \tau^*(x))$ we write $e_{\tau^*}$ resp. $D_{\tau^*}$.

**Lemma 2.6.** *Given a policy $u \in \mathcal{U}$, the corresponding cost functional $J_\lambda^u \in \mathbb{R}^{|\mathcal{S}|}$ is given by*

$$J_\lambda^u = (I - D_\tau P_u)^{-1}(C_u + k_{\text{info}} e_\tau), \tag{2.16}$$

*where $C_u(x) := C(x, a(x), \tau(x))$ for all $x \in \mathcal{S}$ and $u(x) = (a(x), \tau(x))$.*

*Proof.* Recursion (2.9) can be written in the form

$$J^u_\lambda = C_u + k_{\text{info}} e_\tau + D_\tau P_u J^u_\lambda,$$

which is equivalent to

$$(I - D_\tau P_u) J^u_\lambda = C_u + k_{\text{info}} e_\tau.$$

Now the matrix $I - D_\tau P_u$ is invertible by the following argumentation. If it was not invertible, the equation

$$(I - D_\tau P_u) v = 0$$

would have a solution $v \in \mathbb{R}^{|\mathcal{S}|} \neq 0$. We write $v = D_\tau P_u v$ and note that for those states $x \in \mathcal{S}$ with $\tau(x) = \infty$ it holds $D_\tau(x, y) = 0$ for all $y \in \mathcal{S}$ which delivers $v(x) = 0$. Hence, we can assume $\tau(x) < \infty$ for all $x \in \mathcal{S}$; otherwise we consider only those components of the equation referring to states with finite lag time. Now, as $D_\tau$ is a diagonal matrix with diagonal entries $0 < e^{-\lambda\tau(x)} < 1$, its inverse $D_\tau^{-1}$ exists and is again diagonal with $D_\tau^{-1}(x, x) = e^{\lambda\tau(x)} > 1$. We write $P_u v = D_\tau^{-1} v$ and take the maximum norm on both sides (which is finite because $\mathcal{S}$ is finite). As $P_u$ is a transition matrix, the entries of $P_u v$ are convex combinations of the entries of $v$, such that $||P_u v||_\infty \leq ||v||_\infty$ holds. On the other hand, it holds that $||D_\tau^{-1} v||_\infty > ||v||_\infty$, as each entry of $v$ is multiplied by a constant larger than 1. Together we get

$$||v||_\infty \geq ||P_u v||_\infty = ||D_\tau^{-1} v||_\infty > ||v||_\infty,$$

a contradiction to $v \neq 0$.
Taking the inverse of $I - D_\tau P_u$ completes the proof. □

The properties of the operator $T_u$ considered in Section 1.1.1 are also valid for the operator $T_u$ defined in (2.11). Along the lines of Lemma 1.7, we formulate

**Lemma 2.7.** *If a function $J$ on $\mathcal{S}$ fulfills*

$$J(x) \geq (T_u J)(x) \quad \forall x \in \mathcal{S}, \tag{2.17}$$

*then $J(x) \geq J_\lambda(x, u)$ $\forall x \in \mathcal{S}$.*

*Proof.* Analogous to the proof of Lemma 1.7. □

Again, it is evident that $T_u$ is a monotone operator, such that the sequence of functions defined by

$$J_0 := 0, \quad J_{n+1} := T_u J_n \quad \text{for } n \geq 0$$

is non-decreasing with $J_n(x) \overset{n\to\infty}{\longrightarrow} J_\lambda(x, u)$ for all $x \in \mathcal{S}$.

### Optimal policy and value function

The preparatory work enables to directly reproduce the results for optimal policies from Section 1.1.1 in the new setting and the value function $V_\lambda(x) = \inf_{u\in\mathcal{U}} J_\lambda(x,u)$ defined in (2.4).

**Theorem 2.8** (Discounted cost optimality equation/Bellman equation)**.** *The value function $V_\lambda$ satisfies the recursion*

$$V_\lambda(x) = \inf_{a\in\mathcal{A},\tau\in(0,\infty]} \left\{ C(x,a,\tau) + e^{-\lambda\tau}\left(k_{\text{info}} + \sum_{y\in\mathcal{S}} P_{a,\tau}(x,y)V_\lambda(y)\right)\right\} \quad \forall x\in\mathcal{S}. \tag{2.18}$$

*Proof.* Analogous to the proof of Theorem 1.8: Equation (2.18) can be written as $V_\lambda(x) = \inf_{a\in\mathcal{A},\tau\in(0,\infty]}\left(T_{a,\tau}V_\lambda\right)(x)$. For an arbitrary $u\in\mathcal{U}$ it holds that $J^u_\lambda(x) \geq V_\lambda(x)$ $\forall x\in\mathcal{S}$ and, as $T_{a,\tau}$ is a monotone operator for each $a\in\mathcal{A}$ and $\tau\in(0,\infty]$, we get

$$J^u_\lambda(x) = (T_{u(x)}J^u_\lambda)(x) \geq (T_{u(x)}V_\lambda)(x) \geq \inf_{a\in\mathcal{A},\tau\in(0,\infty]}(T_{a,\tau}V_\lambda)(x).$$

This implies

$$V_\lambda(x) = \inf_{u\in\mathcal{U}} J^u_\lambda(x) \geq \inf_{a\in\mathcal{A},\tau\in(0,\infty]}(T_{a,\tau}V_\lambda)(x).$$

On the other hand, assuming $V_\lambda(x^*) > \inf_{a\in\mathcal{A},\tau\in(0,\infty]}(T_{a,\tau}V_\lambda)(x^*)$ for an $x^*\in\mathcal{S}$, yields

$$V_\lambda(x) = \inf_{a\in\mathcal{A},\tau\in(0,\infty]}(T_{a,\tau}V_\lambda)(x) + \varepsilon(x) \quad \forall x\in\mathcal{S},$$

where $\varepsilon(x)\geq 0$ for all $x\in\mathcal{S}$ and $\varepsilon(x^*)>0$. Depending on $x$ we can choose an action $a(x)$ and a lag time $\tau(x)\in(0,\infty]$ such that

$$\begin{aligned} V_\lambda(x) &\geq (T_{a(x),\tau(x)}V_\lambda)(x) + \frac{\varepsilon(x)}{2} \\ &= C(x,a,\tau) + \frac{\varepsilon(x)}{2} + e^{-\lambda\tau}\left(k_{\text{info}} + \sum_{y\in\mathcal{S}} P_{a,\tau}(x,y)V_\lambda(y)\right) \\ &=: (\tilde{T}_u V_\lambda)(x), \end{aligned}$$

where $\tilde{T}_u$ is an operator referring to a new cost function $\tilde{C}(x,a,\tau) := C(x,a,\tau) + \frac{\varepsilon(x)}{2}$. By $\tilde{C}(x^*,a,\tau) > C(x^*,a,\tau)$ we get

$$\begin{aligned} \tilde{J}_\lambda(x^*,u) &= \mathbb{E}^u_{x^*}\left(\sum_{\substack{j\in\mathbb{N}_0\\ t_j<\infty}} e^{-\lambda t_j}\left(\tilde{C}(X_{t_j},a(X_{t_j}),\tau(X_{t_j})) + e^{-\lambda\tau(X_{t_j})}k_{\text{info}}\right)\right) \\ &> J_\lambda(x^*,u) \end{aligned}$$

for the cost functional determined by $\tilde{C}$ and $\tilde{T}_u$. By Lemma 2.7 we can deduce $V_\lambda(x^*) \geq \tilde{J}_\lambda(x^*,u) > J_\lambda(x^*,u)$, in contradiction to $V_\lambda(x^*) = \inf_{u\in\mathcal{U}} J_\lambda(x^*,u)$. □

Equation (2.18) characterizes not only the value function $V_\lambda$, but also the optimal policy, as long as the infimum can be replaced by a minimum which will be assumed in the following theorem.

**Theorem 2.9.** *Suppose that there exists a policy $u^*(x) = (a^*(x), \tau^*(x)) \in \mathcal{U}$ which attains the minimum in the Bellman equation (2.18), i.e.*

$$\begin{aligned} V_\lambda(x) &= C(x, a^*(x), \tau^*(x)) + e^{-\lambda\tau^*(x)} \left( k_{\text{info}} + \sum_{y\in\mathcal{S}} P_{a^*(x),\tau^*(x)}(x,y) V_\lambda(y) \right) \quad (2.19) \\ &= \min_{a\in\mathcal{A},\tau\in(0,\infty]} \left\{ C(x,a,\tau) + e^{-\lambda\tau} \left( k_{\text{info}} + \sum_{y\in\mathcal{S}} P_{a,\tau}(x,y) V_\lambda(y) \right) \right\} \quad \forall x \in \mathcal{S}. \end{aligned}$$

*Then $u^*$ is discounted-cost optimal, i.e. it holds $V_\lambda(x) = J_\lambda(x, u^*)$ for all $x \in \mathcal{S}$.*

*Proof.* Analogous to the proof of Theorem 1.10, using of Lemma 2.7 instead of Lemma 1.7. □

Let us again (as in Chapter 1) assume $\mathcal{A}$ to be finite. By the lag time parameter $\tau$ the set of policies in our new setting is extended from the discrete set $\mathcal{A}^{\mathcal{S}}$ to the uncountable set $(\mathcal{A} \times (0,\infty])^{\mathcal{S}}$. This means that even for finite state and action spaces the existence of an optimal policy given the cost criterion $J_\lambda(x,u)$ is not self-evident. Fortunately, we can observe that the cost functional $J_\lambda(x,u)$ is a continuous function of the lag times $\tau(x)$ determined by the policy $u$, see Lemma 2.10. Noting further that

$$\lim_{\tau(x)\to 0} J_\lambda(x,u) = \infty \quad \forall x \in \mathcal{S},$$

which is due to $k_{\text{info}} > 0$, we can locate a lower bound $\varepsilon > 0$ for the optimal lag times. The continuity of $J_\lambda(x,u)$ together with the compactness of $[\varepsilon,\infty]$ guarantees – for finite $\mathcal{A}$ – the existence of a policy $u^* \in \mathcal{U}$ with $J_\lambda(x,u^*) = \min_{u\in\mathcal{U}} J_\lambda(x,u)$.

**Lemma 2.10** (Continuity of $J_\lambda(x,u)$ with respect to $\tau(x)$)**.** *For all $x \in \mathcal{S}$ and $a \in \mathcal{A}$, the cost functional $J_\lambda(x,u)$ of a given policy $u(x) = (a(x), \tau(x))$ is continuous with respect to the lag time parameter $\tau(x) \in (0,\infty]$.*

*Proof.* The continuity of $J_\lambda(x,u)$ with respect to $\tau = \tau(x)$ follows from the continuity of the expressions $e^{-\lambda\tau}$, $e^{L_a\tau}$ and $C(x,a,\tau)$ which are the $\tau$-dependent components in

$$J_\lambda^u = (I - D_\tau P_u)^{-1}(C_u + k_{\text{info}} e_\tau),$$

compare Lemma 2.6 and the definitions given in (2.12) and (2.14). Especially, the continuity at $\tau(x) = \infty$ follows from the definitions given in (2.2). □

Note that, instead of assuming $\mathcal{A}$ to be finite in order to guarantee the existence of an optimal policy, we can use Assumption 1.11 or any other constraints which ensure the existence of an optimal policy in the original setting.

### Numerical realization

From

$$(T^*J)(x) := \min_{a\in\mathcal{A},\tau\in(0,\infty]} (T_{a,\tau}J)(x)$$

and $V_0 = 0$, $V_{n+1} = T^*V_n$ $(n \in \mathbb{N})$ we get a nondecreasing sequence of functions converging to the value function $V_\lambda$ and defining implicitly a sequence of policies converging to the optimal policy. The justifying argumentation is analogous to the proof of Theorem 1.12. This directly delivers a value iteration approach to numerically determine the optimal policy of the considered Markov control problem with information costs.
Likewise, the policy iteration algorithm described in Section 1.2 can be transferred to the new framework which yields

**The discounted-cost policy iteration algorithm:**

1. Pick an arbitrary $u \in \mathcal{U}$. Let $k = 0$ and $u_0 := u$.

2. (Policy evaluation) Obtain $J_\lambda^k = (I - D_\tau P_u)^{-1}(C_u + k_{\text{info}}e_\tau)$.

3. (Policy improvement) For each $x \in \mathcal{S}$ calculate

$$m(x) := \min_{a\in\mathcal{A},\tau\in(0,\infty]} \left(T_{a,\tau}J_\lambda^k\right)(x)$$

and set $(a_k(x), \tau_k(x)) := \arg\min_{a\in\mathcal{A},\tau\in(0,\infty]} \left(T_{a,\tau}J_\lambda^k\right)(x)$.
Define $u_{k+1} = (a_{k+1}, \tau_{k+1})$ as follows:

$$u_{k+1}(x) := \begin{cases} (a_k(x), \tau_k(x)), & \text{if } m(x) < J_\lambda^k(x), \\ u_k(x), & \text{if } m(x) = J_\lambda^k(x). \end{cases}$$

4. If $u_{k+1} = u_k$, then stop. Otherwise increment $k$ by 1 and return to Step 2.

It should be noticed that the policy improvement step now contains a minimization with respect to the continuous parameter $\tau \in (0, \infty]$ of lag times. One approach to numerically perform this step is to simply discretize the domain $(0, \infty]$ of the lag time parameter $\tau$. In many real-world applications, the lag time parameter naturally exhibits some kind of discrete quality given by the considered time unit (years/days/hours/seconds ...). In such a case, an arbitrarily precise placing of interaction times might not be possible, see for example the application described in Chapter 3. This naturally suggests to consider discrete lag times. Such a discretization depending on the given problem strongly simplifies the policy improvement step. All numerical calculations within this work are based on such a discretization and always showed a satisfying performance.
Still, the question is whether there exist other numerical approaches (other than policy iteration) which are suited to determine the optimal policy within the new setting of Markov control with information costs. The optimization problem consists of minimizing the cost functional $J_\lambda^u$ over the domain of policies given by

$(\mathcal{A}\times(0,\infty])^{\mathcal{S}}$. As we consider finite sets $\mathcal{A}$ of actions, this domain has both discrete and continuous components. However, all types of continuous optimization methods (like the gradient descent method, Newton's method or convex programming) require the underlying domain over which they optimize to be connected subsets of $\mathbb{R}^n$ ($n \in \mathbb{N}$) [3,8,15]. These methods assume the objective function to be differentiable and/or convex on the given domain which precludes from considering discrete variables. For methods of discrete optimization, on the other hand, it is not clear how to deal with the continuous lag time parameter $\tau \in (0,\infty]$.
This means that these methods are not able to *replace* the policy iteration algorithm. Instead, they can be *combined* with the policy iteration approach by applying them in each policy improvement step. For example, one could use the gradient descent method in order to determine the minimum with respect to the lag time parameter $\tau$ for fixed $a \in \mathcal{A}$. However, as we will see in Section 2.2.4, the sensitivity of the cost functional with respect to the lag time parameter can be very low which would imply a slow convergence of the gradient descent method.
In summary, we can state that – as long as state and action space are not too large – the presented policy iteration combined with a discretization of the lag time parameter is totally suited to numerically calculate the optimal control policy for a Markov decision process with information costs. In this regard, also multi-level approaches could help to efficiently calculate the optimal lag time $\tau$ for fixed $a \in \mathcal{A}$ by combining different discretization levels for the set $(0,\infty]$.

Several examples will illustrate our results.

**Example 2.11** (Two states)**.** *Once again consider the 2-state-example 1.2 introduced on page 9 with $\mathcal{S} = \{x_1, x_2\}$, $\mathcal{A} = \{a_1, a_2\}$,*

$$L_1 = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix}, \quad L_2 = \begin{pmatrix} -0.1 & 0.1 \\ 0.1 & -0.1 \end{pmatrix}$$

*and $c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$, where $c_{\mathcal{S}}(x_1) = 0$, $c_{\mathcal{S}}(x_2) > 0$, $c_{\mathcal{A}}(a_1) = 0$, $c_{\mathcal{A}}(a_2) > 0$. The optimal policy in the original setting (i.e. assuming full observability of the process) for $c_{\mathcal{S}}(x_2) = 10$ and $c_{\mathcal{A}}(a_2) = 2$ was given by $u^*(x_1) = a_1$ and $u^*(x_2) = a_2$, resulting in the value function $V_\lambda(x_1) = 5.71$, $V_\lambda(x_2) = 62.86$, compare Table 1.1. Now we assume that for each observation of the process we have to pay a fee $k_{\text{info}} > 0$. One can expect the optimal policy to have the following form: Being in state $x_1$ one chooses action $a_1$ for some time period $\tau(x_1)$. Then a test determines the actual state. If this state is $x_2$ one chooses action $a_2$ for another time period $\tau(x_2)$ in order to quickly return to state $x_1$. As action $a_2$ is more expensive and the dynamics given generator $L_2$ are faster, it should hold $\tau(x_1) > \tau(x_2)$.*
*We calculate the optimal policy and the corresponding value function for different values of $c_{\mathcal{S}}(x_2), c_{\mathcal{A}}(a_2), k_{\text{info}}, \lambda$ by the presented policy iteration algorithm. The results are given in Table 2.1.*

*One can observe that increasing the information costs $k_{\text{info}}$ leads to longer optimal lag times for both states (compare the first two rows), while increasing the action costs $c_{\mathcal{A}}(a_2)$ results in a larger value of $\tau^*(x_1)$ combined with a smaller value of*

| $c_{\mathcal{S}}(x_2)$ | $c_{\mathcal{A}}(a_2)$ | $k_{\text{info}}$ | $\lambda$ | $a^*(x_1)$ | $a^*(x_2)$ | $\tau^*(x_1)$ | $\tau^*(x_2)$ | $V_\lambda(x_1)$ | $V_\lambda(x_2)$ |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 2 | 1 | 0.1 | $a_1$ | $a_2$ | 11.3 | 1.8 | 7.78 | 69.77 |
| 10 | 2 | 2 | 0.1 | $a_1$ | $a_2$ | 19.7 | 2.6 | 8.2 | 72.3 |
| 10 | 3 | 1 | 0.1 | $a_1$ | $a_2$ | 13.7 | 1.6 | 8.0 | 75.5 |
| 5 | 2 | 1 | 0.1 | $a_1$ | $a_2$ | 46.6 | 2.1 | 4.2 | 42.0 |
| 10 | 2 | 1 | 0.5 | $a_1$ | $a_1$ | $\infty$ | $\infty$ | 0.4 | 19.6 |
| 10 | 2 | 10 | 0.1 | $a_1$ | $a_2$ | $\infty$ | $\infty$ | 8.3 | 86.7 |

Table 2.1: **Parameter dependent optimal policy and value function.** This table shows the optimal policy for the 2-state-example 2.11 and different values of $c_{\mathcal{S}}(x_2)$, $c_{\mathcal{A}}(a_2)$, $k_{\text{info}}$ and $\lambda$, given the discounted-cost criterion. In state $x_1$ (resp. $x_2$) one has to choose action $a^*(x_1)$ (resp. $a^*(x_2)$) for a time period $\tau^*(x_1)$ (resp. $\tau^*(x_2)$) which results in the value function $V_\lambda$ given by $V_\lambda(x_1)$, $V_\lambda(x_2)$. Advice: Compare each row to the first one.

$\tau^*(x_2)$ *(compare rows 1 and 3). In both cases the value function increases for both states. Decreasing the state costs* $c_{\mathcal{S}}(x_2)$ *also increases the optimal lag times but decreases the value function (compare row 4 to row 1). For a "high" discount factor* $\lambda = 0.5$*, state testing is not profitable at all* ($\tau^*(x_1) = \tau^*(x_2) = \infty$*, see row 5). Instead, the strong discount of future costs eases the damage caused by being in state* $x_2$ *such that a quick leaving is not necessary anymore* ($a^*(x_2) = a_1$)*. A complete omission of state testing is also caused by high information costs* $k_{\text{info}} = 10$*, however in this case we get* $a^*(x_2) = a_2$ *(see row 6).*

In the following, we extend the 2-state-model by adding an intermediate state $x_I$ lying "in between" the two state $x_1$ and $x_2$, compare Figure 2.2.

**Example 2.12** (Three states)**.** *We take the 2-state-model (Example 2.11) and add an intermediate state* $x_I$ *which represents the transition area. This state* $x_I$ *gets the same cost value as state* $x_1$*, i.e.* $c_{\mathcal{S}}(x_I) = c_{\mathcal{S}}(x_1) = 0$*, but higher exit rates. When interpreting* $x_1$ *as the "good" state and* $x_2$ *as the "bad" state, then the new state* $x_I$ *is still "good", but transitions from* $x_I$ *to the "bad" state* $x_2$ *become pretty likely. Finding the process in state* $x_I$ *stands for ringing the alarm bell: We need to keep the process from switching into the "bad"/costly state* $x_2$ *where switching back to* $x_I$ *becomes difficult. The goal is to compare the 3-state-model to the 2-state-model. We would expect an improvement of the values (i.e. lower costs for the optimal policy) because of the warning character of the intermediate state.*

*Given the original generators* $L_1$*,* $L_2$ *of the 2-state-model and some new transition rates* $r_a(x_1, x_I)$*,* $r_a(x_I, x_1)$ *between the states* $x_1$ *and* $x_I$ *(which indicate how much the states* $x_I$ *and* $x_1$ *are connected), we want to find a transition rate* $r_a(x_I, x_2)$*, between* $x_I$ *and* $x_2$ *such that the equilibrium distribution in* $x_2$*, comparing the new 3-state-model to the original 2-state-model, stays the same. It is easy to check that one has to choose*

$$r_a(x_I, x_2) = l_a(x_1, x_2) \cdot \frac{r_a(x_1, x_I) + r_a(x_I, x_1)}{r_a(x_1, x_I)}.$$

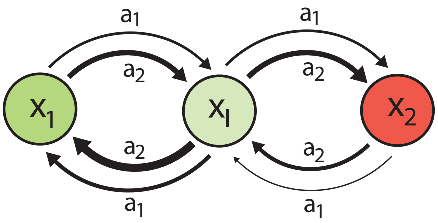

Figure 2.2: **3-state-example.** Action $a_2$ increases the transition rates between the states $x_1$, $x_I$ and $x_2$, which is indicated by the thickness of the corresponding arrows.

*The new generators are then given by*

$$\tilde{L}_a = \begin{pmatrix} -r_a(x_1,x_I) & r_a(x_1,x_I) & 0 \\ r_a(x_I,x_1) & -r_a(x_I,x_1) - r_a(x_I,x_2) & r_a(x_I,x_2) \\ 0 & l_a(x_2,x_1) & -l_a(x_2,x_1) \end{pmatrix}.$$

*We consider the concrete values from the first line of Table 2.1 in Example 2.11 and calculate the optimal policy of the new 3-state-model for different transition rates between $x_1$ and $x_I$. Choosing $r_1(x_1,x_I) = 0.05$, $r_2(x_1,x_I) = 0.5$, $r_1(x_I,x_1) = 0.1$, $r_2(x_I,x_1) = 1$ we get*

$$\tilde{L}_1 = \begin{pmatrix} -0.05 & 0.05 & 0 \\ 0.1 & -0.13 & 0.03 \\ 0 & 0.01 & -0.01 \end{pmatrix}, \quad \tilde{L}_2 = \begin{pmatrix} -0.5 & 0.5 & 0 \\ 1 & -1.3 & 0.3 \\ 0 & 0.1 & -0.1 \end{pmatrix},$$

*and the optimal policy is given by*

$$a^*(x_1) = a_1, \quad a^*(x_I) = a_1, \quad a^*(x_2) = a_2,$$
$$\tau^*(x_1) = 17.8, \quad \tau^*(x_I) = 6.4, \quad \tau^*(x_2) = 1.8,$$

*with the value function*

$$V_\lambda(x_1) = 4.5, \quad V_\lambda(x_I) = 12.9, \quad V_\lambda(x_2) = 72.6.$$

*Comparing this to the original values of the 2-state-model ($\tau^*(x_1) = 11.3$, $\tau^*(x_2) = 1.8$, $V_\lambda(x_1) = 7.78$, $V_\lambda(x_2) = 69.77$), we can see that for state $x_1$ the optimal time $\tau$ grows and the value function decreases, while for state $x_2$ the time stays the same and the value function increases. The values for new state $x_I$ are arranged in between. The increase in $V_\lambda(x_2)$ is due to the fact that when leaving state $x_2$ the process does not directly enter the (more or less) "safe" state $x_1$. Instead, it reaches the transition area $x_I$ where a return to $x_2$ is more likely.*

*By increasing the transition rates $r_a$ between $x_1$ and $x_I$, the results converge to the original numbers: Choosing for example $r_1(x_1,x_I) = 100$, $r_2(x_1,x_I) = 1000$, $r_1(x_I,x_1) = 100$, $r_2(x_I,x_1) = 1000$ results in*

$$\tilde{L}_1 = \begin{pmatrix} -100 & 100 & 0 \\ 100 & -100.02 & 0.02 \\ 0 & 0.01 & -0.01 \end{pmatrix}, \quad \tilde{L}_2 = \begin{pmatrix} -1000 & 1000 & 0 \\ 1000 & -1000.3 & 0.3 \\ 0 & 0.1 & -0.1 \end{pmatrix},$$

*and the optimal policy is given by*

$$a^*(x_1) = 1, \quad a^*(x_I) = 1, \quad a^*(x_2) = 2,$$

$$\tau^*(x_1) = 11.3, \quad \tau^*(x_I) = 11.3, \quad \tau^*(x_2) = 1.8,$$

*with the value function*

$$V_\lambda(x_1) = 7.8, \quad V_\lambda(x_I) = 7.8, \quad V_\lambda(x_2) = 69.8.$$

*The state $x_I$ looses its role of an intermediate state and merges with state $x_1$.*

In order to learn more about the possible structure of the optimal policy and the value function, we reconsider the population process that was introduced in Section 1.2 as a third example .

**Example 2.13** (Controlled Population (cont.)). *We consider again the controlled birth-death-process described in Example 1.18, i.e. the state of the process is given by the number of individuals and the goal is to decrease the population size. This time, we assume that the process cannot be observed for free. Instead, each observation produces costs $k_{\text{info}} = 1$. Let the other parameters be the same as in Example 1.18. We calculate the optimal policy for two kinds of cost functions $c_{\mathcal{S}}$: linear costs $c_{\mathcal{S}}(x) = x$ and quadratic costs $c_{\mathcal{S}}(x) = x^2$. The results are shown in Figure 2.3 and 2.4.*

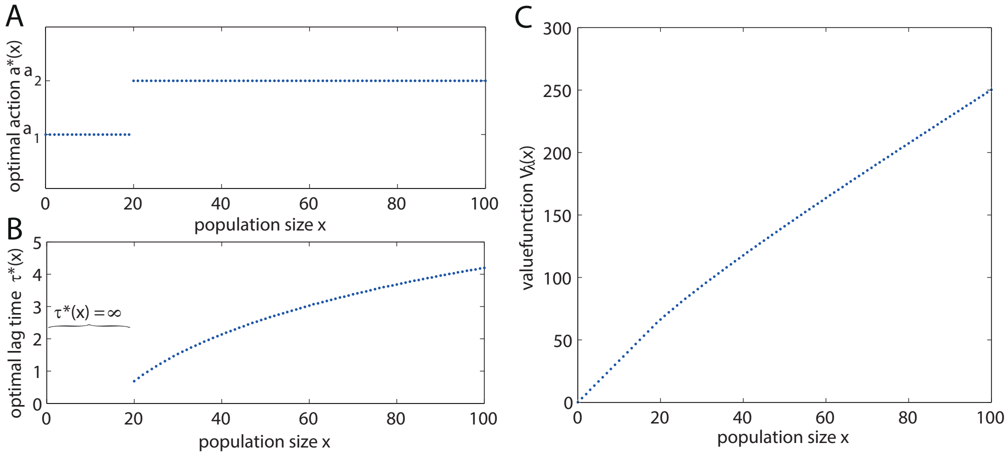

Figure 2.3: **Optimal policy for controlled population and $c_{\mathcal{S}}(x) = x$.** Given the information that the process from Example 2.13 is at time $t$ in state $x$, one has to choose action $a^*(x)$ (shown in panel A) for a time period $\tau^*(x)$ (shown panel B). The next information is taken at time $t + \tau^*(x)$. The related optimal discounted costs are given by $V_\lambda(x)$ (shown in panel C). For $x \in \{0, ..., 19\}$ we get $\tau^*(x) = \infty$.

*As one can see – and as one would expect –, the optimal policy is in both cases of the form $a^*(x) = a_1 \ \forall x \leq x^*$ and $a^*(x) = a_2 \ \forall x > x^*$ for some critical state $x^* \in \mathcal{S}$. That is, the more expensive action $a_2$ which increases the death rate, is* *applied whenever the population size exceeds a fixed number. For quadratic costs ($x^* = 3$) this critical state is smaller than for linear costs ($x^* = 19$). While the optimal lag times $\tau^*(x)$ are monotonously increasing in $x$ for states $x > x^*$, they are infinite (given linear costs) resp. decreasing (given quadratic costs) for states $x \leq x^*$.*

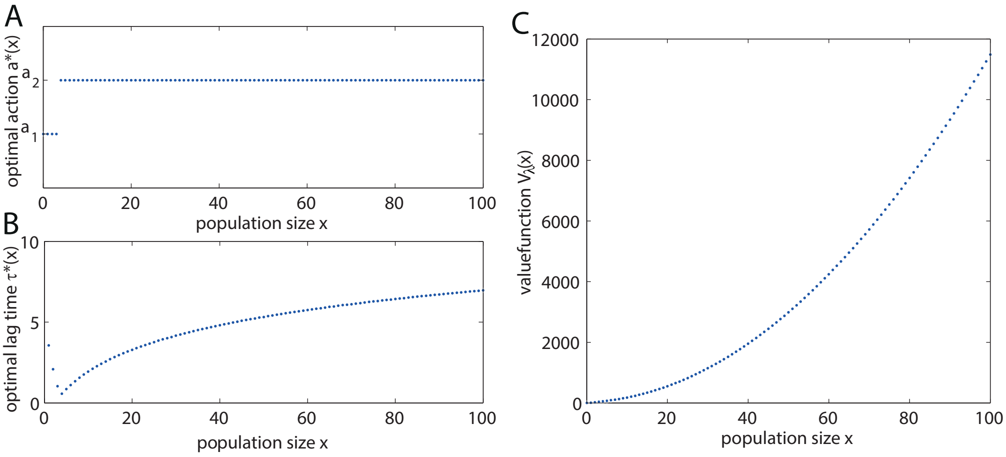

Figure 2.4: **Optimal policy for controlled population and $c_S(x) = x^2$.** Given the information that the process in Example 2.13 is at time $t$ in state $x$, one has to choose action $a^*(x)$ (shown in panel A) for a time period $\tau^*(x)$ (shown in panel B). The next information is taken at time $t + \tau^*(x)$. The related optimal discounted costs are given by $V_\lambda(x)$ (shown in panel C). It is $\tau^*(0) = \infty$.

### 2.2.2 Cost Splitting for Discounted Costs

Given the Markov control model with information costs (2.1), an evident question is how the corresponding value function is related to the value function of the original control problem (with complete observability). In the new setting, the value function contains not only the process costs (defined by the cost function $c$) but also the information costs induced by $k_{\text{info}}$ under optimal control. Therefore, in order to carry out a comparison, we initially have to calculate the *net costs* for the new setting, which are the total expected discounted costs without the information costs:

$$J_{\text{net}}(x, u) := \mathbb{E}_x^u \left( \sum_{\substack{j \in \mathbb{N}_0 \\ t_j < \infty}} \int_{t_j}^{t_{j+1}} e^{-\lambda s} c(X_s, a(X_{t_j}))\, ds \right).$$

Especially, we are interested in the net costs under optimal control $u^*$,

$$V_{\text{net}}(x) := J_{\text{net}}(x, u^*).$$

In line with this, the *information costs* under optimal control are given by

$$V_{\text{info}}(x) := \mathbb{E}_x^{u^*}\left(\sum_{\substack{j\in\mathbb{N}_0\\ t_j<\infty}} e^{-\lambda t_{j+1}} k_{\text{info}}\right). \tag{2.20}$$

Remember that the observation times $(t_j)_{j\in\mathbb{N}_0}$ are random variables with a distribution determined by the initial state $x \in \mathcal{S}$ and the policy $u^*$.
Now we can split up the value function $V_\lambda$ into two parts:

$$V_\lambda = V_{\text{net}} + V_{\text{info}}.$$

The goal is to calculate this function $V_{\text{net}}$ by means of the value function $V_\lambda$ and the optimal policy $u^*$.

A decomposition of the sum in equation (2.20) directly yields the recursion

$$V_{\text{info}}(x) = e^{-\lambda\tau^*(x)} k_{\text{info}} + e^{-\lambda\tau^*(x)} \sum_{y\in\mathcal{S}} P_{u^*}(x,y) V_{\text{info}}(y). \tag{2.21}$$

From this we can deduce the following

**Lemma 2.14** (Information costs)**.** *For the information costs under optimal control it holds*

$$V_{\text{info}} = k_{\text{info}}(I - D_{\tau^*}P_{u^*})^{-1} e_{\tau^*},$$

*where $I \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ is the identity matrix.*

*Proof.* Equivalent to the proof of Lemma 2.6. □

That is, once we know the optimal policy, we can calculate the information costs $V_{\text{info}}$ and the net costs $V_{\text{net}}$.
A further splitting up of the value function is possible if the cost function $c$ is of the form

$$c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a),$$

which was the case in several of the considered examples. This means that the costs induced by the state are independent of the costs induced by the chosen action. It suggests itself to define the *action costs* under optimal control $u^*(x) = (a^*(x), \tau^*(x))$ by

$$V_{\mathcal{A}}(x) := \mathbb{E}_x^{u^*}\left(\sum_{\substack{j\in\mathbb{N}_0\\ t_j<\infty}} \int_{t_j}^{t_{j+1}} e^{-\lambda s} c_{\mathcal{A}}(a^*(X_{t_j}))\, ds\right)$$

and the *state costs* under optimal control $u^*$ by

$$V_{\mathcal{S}}(x) := \mathbb{E}_x^{u^*}\left(\sum_{\substack{j\in\mathbb{N}_0\\ t_j<\infty}} \int_{t_j}^{t_{j+1}} e^{-\lambda s} c_{\mathcal{S}}(X_s)\, ds\right).$$

Given this notation, it is $V_{\text{net}} = V_{\mathcal{S}} + V_{\mathcal{A}}$ as well as

$$V_\lambda = V_{\mathcal{S}} + V_{\mathcal{A}} + V_{\text{info}}.$$

For the function $V_{\mathcal{A}}$ of optimal action costs it holds

$$\begin{aligned} V_{\mathcal{A}}(x) &= \int_0^{\tau^*(x)} e^{-\lambda s} c_{\mathcal{A}}(a^*(x))\, ds + e^{-\lambda\tau^*(x)} \sum_{y\in\mathcal{S}} P_{u^*}(x,y) V_{\mathcal{A}}(y) \\ &= \frac{1}{\lambda}(1 - e^{-\lambda\tau^*(x)}) c_{\mathcal{A}}(a^*(x)) + e^{-\lambda\tau^*(x)} \sum_{y\in\mathcal{S}} P_{u^*}(x,y) V_{\mathcal{A}}(y), \end{aligned} \tag{2.22}$$

which results in

**Lemma 2.15** (Action costs)**.** *For the action costs $V_{\mathcal{A}}$ under optimal control it holds*

$$V_{\mathcal{A}} = \frac{1}{\lambda}(I - D_{\tau^*}P_{u^*})^{-1}(I - D_{\tau^*})c^*_{\mathcal{A}},$$

*where $c^*_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{S}|}$ is defined by $c^*_{\mathcal{A}}(x) = c_{\mathcal{A}}(a^*(x))$ for all $x \in \mathcal{S}$.*

*Proof.* The recursion (2.22) can be written as

$$V_{\mathcal{A}} = \frac{1}{\lambda}(I - D_{\tau^*})c^*_{\mathcal{A}} + D_{\tau^*}P_{u^*}V_{\mathcal{A}}.$$

This directly leads to

$$(I - D_{\tau^*}P_{u^*})V_{\mathcal{A}} = \frac{1}{\lambda}(I - D_{\tau^*})c^*_{\mathcal{A}}$$

and

$$V_{\mathcal{A}} = \frac{1}{\lambda}(I - D_{\tau^*}P_{u^*})^{-1}(I - D_{\tau^*})c^*_{\mathcal{A}}.$$

The fact that the matrix $I - D_{\tau^*}P_{u^*}$ is invertible was shown in the proof of Lemma 2.6. □

Having deduced an analytical formula for both the information costs and the action costs under optimal control, we want to do the same for the optimal state costs $V_{\mathcal{S}}$. By a decomposition we get

$$V_{\mathcal{S}}(x) = \mathbb{E}_x^{u^*}\left(\int_0^{\tau^*(x)} e^{-\lambda s} c_{\mathcal{S}}(X_s)\, ds\right) + e^{-\lambda\tau^*(x)} \sum_{y\in S} P_{u^*}(x,y) V_{\mathcal{S}}(y), \tag{2.23}$$

and we get

**Lemma 2.16** (State costs)**.** *For the state costs $V_{\mathcal{S}}$ under optimal control it holds*

$$V_{\mathcal{S}} = (I - D_{\tau^*}P_{u^*})^{-1} H\, c_{\mathcal{S}}$$

*with $H \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ defined by*

$$H(x,y) := \left(L_{a^*(x)} - \lambda I\right)^{-1} \left(e^{(L_{a^*(x)} - \lambda I)\tau^*(x)} - I\right)(x,y), \quad x, y \in \mathcal{S}.$$

*Proof.* For the first summand in (2.23) we have

$$\begin{aligned}
\mathbb{E}_x^{u^*}\left(\int_0^{\tau^*(x)} e^{-\lambda s} c_{\mathcal{S}}(X_s)\, ds\right) &= \int_0^{\tau^*(x)} e^{-\lambda s} \left(e^{L_{a^*(x)} s} c_{\mathcal{S}}\right)(x)\, ds \\
&= \int_0^{\tau(x)} \left(e^{(L_{a^*(x)}-\lambda I)s} c_{\mathcal{S}}\right)(x)\, ds \\
&= \left[(L_{a^*(x)} - \lambda I)^{-1}\left(e^{(L_{a^*(x)}-\lambda I)\tau^*(x)} - I\right) c_{\mathcal{S}}\right](x).
\end{aligned}$$

By the definition of the matrix $H$ this means

$$V_{\mathcal{S}} = H\, c_{\mathcal{S}} + D_{\tau^*} P_{u^*} V_{\mathcal{S}},$$

and

$$(I - D_{\tau^*} P_{u^*})\, V_{\mathcal{S}} = H\, c_{\mathcal{S}}.$$

Taking the inverse of $I - D_{\tau^*} P_{u^*}$ (compare again the proof of Lemma 2.6) proves the statement. □

For the sake of completeness we note a compatible analytic expression for the net costs $V_{\text{net}}$ under optimal control.

**Corollary 2.17** (Net costs)**.** *For the net costs* $V_{\text{net}} = V_{\mathcal{S}} + V_{\mathcal{A}}$ *it holds*

$$V_{\text{net}} = (I - D_{\tau^*} P_{u^*})^{-1} C_{u^*}.$$

Figure 2.5 illustrates the cost splitting for the controlled population from Example 2.13 with linear costs $c_{\mathcal{S}}(x) = x$.

Throughout this section we considered the optimal policy $u^*$ and formulated the cost splitting for the corresponding optimal costs. Of course, such a cost splitting can also be done for any non optimal policy $u \in \mathcal{U}$ with respect to the function $J_\lambda^u$ of total discounted costs under this policy. The analysis is completely analogous and delivers the same analytic expressions for all parts of the cost splitting.

The presented cost splitting is by itself an interesting tool to analyze the structure of the value function for a given control problem. For instance, within the medical application presented in Chapter 3 the state costs will be associated with the health damage of a patient. In this case, it is of fundamental interest to extract the state costs from the total costs in order to assess the impact of a medical therapy on the health status of the patient.
As another advantage, we are now able to make an unbiased comparison to the case of cost-free information, which will be done in the following.

### Comparison to the original Markov control problem

Our intention is to compare the value function of a given Markov control problem with information costs to the value function of the corresponding original Markov control problem. In order to get a first insight, we consider the 2-state-example 2.11

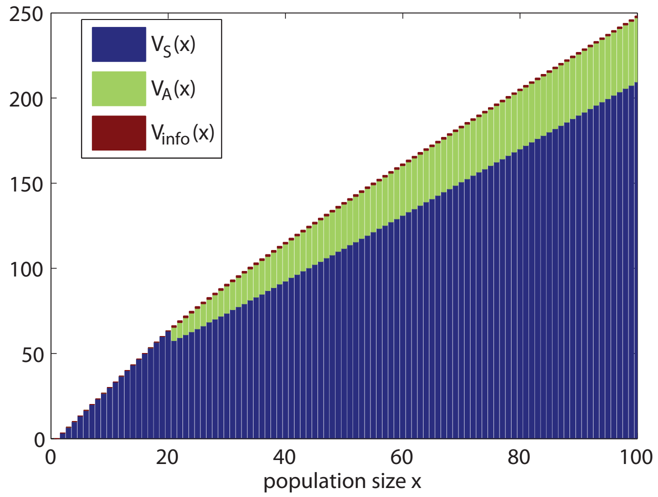

Figure 2.5: **Cost splitting for controlled population and $c_{\mathcal{S}}(x) = x$.** The value function $V_\lambda$ for Example 2.13 (compare Figure 2.3) is divided into its components $V_{\mathcal{S}}$ (blue), $V_{\mathcal{A}}$ (green) and $V_{\text{info}}$ (red). For $x \in \{0, ..., 19\}$ the information and action costs vanish due to the fact that $\tau^*(x) = \infty$ and $a^*(x) = a_1$ with $c_{\mathcal{A}}(a_1) = 0$ for those $x$.

and calculate the cost splitting for the value function of both problems. To this end, we note that in the original control problem the state costs under optimal control $u^*$ are given by

$$(\lambda I - L_{u^*})^{-1} c_{\mathcal{S}},$$

while the action costs under optimal control $u^*$ are given by

$$(\lambda I - L_{u^*})^{-1} c^*_{\mathcal{A}},$$

where $c^*_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{S}|}$ is defined by $c^*_{\mathcal{A}}(x) := c_{\mathcal{A}}(u^*(x))$ (keeping in mind that for the original control problem a deterministic policy is of the form $u : \mathcal{S} \to \mathcal{A}$). The results are given in Figure 2.6.

We can see that for the 2-state-example the optimal net costs ($V_{\text{net}} = V_{\mathcal{S}} + V_{\mathcal{A}}$) of the control problem with information costs clearly exceed the optimal costs of the original control problem which coincides with our intuition: It should be clear that a continuous interaction without information costs can only lead to better results. Nevertheless, we will in the following try to give an explanation of this relation.

First of all note that a policy in the original setting cannot that easily be expressed in terms of the new definition, i.e. by a tuple $(a(x), \tau(x))$. One would have to set $\tau(x) = 0$ for all $x \in \mathcal{S}$ in order to obtain a continuous interaction without time delay; this, however, cannot be handled in the new setting. The idea is to find a superior definition for the term “policy” which fits both the original and the information cost model.

This can be done by assigning to each state $x \in \mathcal{S}$ a tuple $(a(x), \mathcal{T}(x))$, where $a(x)$

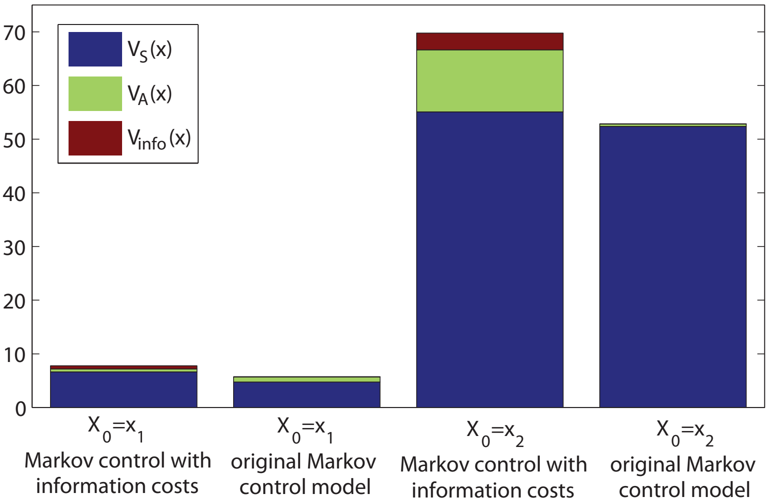

Figure 2.6: **Cost splitting for the 2-state-example.** The value function $V_\lambda$ for Example 2.11 is divided into its components $V_{\mathcal{S}}$ (blue), $V_{\mathcal{A}}$ (green) and $V_{\text{info}}$ (red) and compared to the free information case where the splitting is given by $V_{\mathcal{S}} + V_{\mathcal{A}}$. The respective cost parameters are given by $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_2) = 2$, $k_{\text{info}} = 1$ and $\lambda = 0.1$, compare first row of Table 2.1.

refers, as usual, to an action, while $\mathcal{T}(x)$ is a random variable declaring the length of time until the next adaption of control.
The cost functional (without information costs) for a policy $u(x) = (a(x), \mathcal{T}(x))$ is given by

$$\hat{J}_\lambda(x,u) = \mathbb{E}_x^u \left( \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<\infty}} \int_{t_j}^{t_{j+1}} e^{-\lambda s} c(X_s, a(X_{t_j}))\, ds \right),$$

where $t_0, t_1, \ldots$ are random time points given by $t_0 = 0$ and $t_{j+1} = t_j + \mathcal{T}(X_{t_j})$.

For the Markov control model with information costs, it is $\mathcal{T}(x) = \tau(x)$, i.e. $\mathcal{T}(x)$ is a deterministic function of $x$, and $t_{j+1}$ is $\sigma(t_j, X_{t_j})$-measurable.
For the original model, in contrast, $\mathcal{T}(x)$ is the waiting time in $x$ before a jump occurs, i.e. $\mathcal{T}(x) = \inf_{t>0}\{X_t \neq x | X_0 = x\}$.

Finding an optimal policy in each of the two different settings refers to an optimization within a subset of policies (i.e. an optimization with side constraints). The two sets of admissible policies (policies with deterministic $\mathcal{T}$ and policies with $\mathcal{T}$ given by the jump time of the process) are disjoint. The following consideration shows that for an (overall) optimal policy the times of control adaption have to coincide with the jumping times of the process, i.e. $\mathcal{T}(x)$ has to be the waiting time in $x$ before a jump occurs as in the original control model.

Take an optimal policy $u(x) = (a(x), \mathcal{T}(x))$ and assume that $\mathcal{T}(x)$ is not the

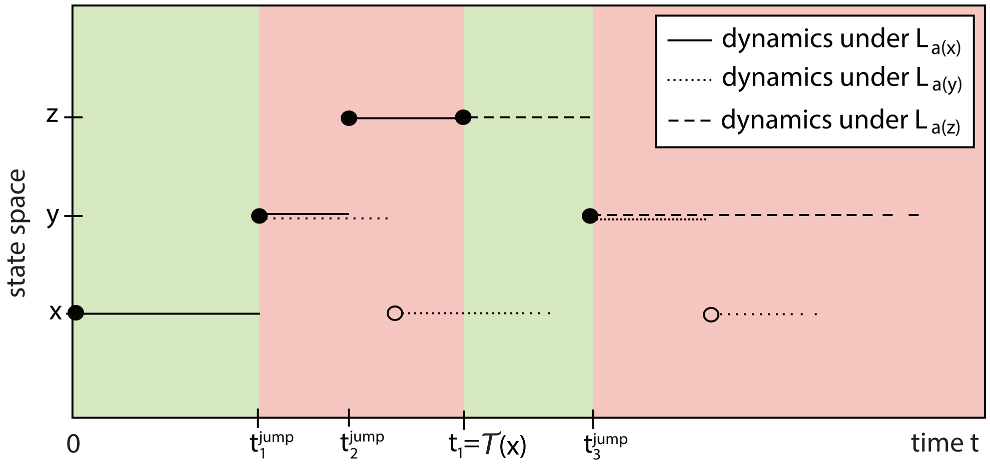

Figure 2.7: **Unadapted control of a Markov process.** Possible trajectory of a controlled process for a policy $u(x) = (a(x), \mathcal{T}(x))$ where $\mathcal{T}(x)$ is *not* the waiting time in $x$. The green areas mark the periods of time where the process control is adapted to the actual state, while the red areas mark the periods of time where the control is not adapted to the actual state and therefore cannot be optimal. The unfilled circles and closely dashed lines refer to a sequel that could arise if the control was continuously adapted.

waiting time in $x$. Imagine the process starting in state $x \in \mathcal{S}$ at time $t = 0$, see Figure 2.7. According to the given optimal policy we chose the generator $L_{a(x)}$ and let the process run until time $\mathcal{T}(x)$. However, with a positive probability, the process switches to another state $y \in \mathcal{S}$ at some point in time $t_1^{\text{jump}} < \mathcal{T}(x)$. We consider this time point $t_1^{\text{jump}}$ as a new starting point of the whole process which is reasonable due to its Markov property. Then the optimal control would prescribe another generator $L_{a(y)}$ (generally $L_{a(y)} \neq L_{a(x)}$). That is, during the time interval $(t_1^{\text{jump}}, \mathcal{T}(x))$ the process is not optimally controlled, which is a contradiction.

From these considerations it immediately follows that the value function of the information cost problem is bounded below by the value function of the original control problem.

Comparing the Markov control problem with information costs to the original control problem suggests to consider the information cost parameter $k_{\text{info}}$ as a variable and to detect how the value function depends on this variable. This will be part of the analysis in the next section.

### 2.2.3 Monotonicity and Continuity with respect to $k_{\text{info}}$

Given a Markov control problem with information costs, we tend to analyze how the information cost parameter $k_{\text{info}}$ influences the value function and the optimal policy. More precisely, we will answer the following questions.

1. Monotonicity of $V_\lambda$ with respect to $k_{\text{info}}$: Given a control problem and changing (only) the cost parameter $k_{\text{info}}$, how does the value function change? Do smaller $k_{\text{info}}$ lead to smaller $V_\lambda(x)$ for all $x \in \mathcal{S}$?

2. Continuity of $V_\lambda$ with respect to $k_{\text{info}}$ at $k_{\text{info}} > 0$: Do small changes in $k_{\text{info}} > 0$ lead to small changes in the value function or are there critical values of $k_{\text{info}}$ where the value function performs a jump?

3. Continuity of $V_\lambda$ with respect to $k_{\text{info}}$ at $k_{\text{info}} = 0$: Considering the limit $k_{\text{info}} \to 0$, does the value function converge (pointwise) to the value function of the original control problem (without information costs)?

4. Monotonicity of $\tau^*$ with respect to $k_{\text{info}}$: Given a control problem and changing (only) the cost parameter $k_{\text{info}}$, how do the optimal lag times $\tau^*(x)$ change? Do smaller $k_{\text{info}}$ lead to smaller $\tau^*(x)$ for all $x \in \mathcal{S}$?

5. Continuity of $\tau^*$ with respect to $k_{\text{info}}$: Considering the limit $k_{\text{info}} \to 0$, do the optimal lag times $\tau^*$ converge (pointwise) to 0 as well?

As we will now show, the answer to each of the first three questions is yes.

### Monotonicity and continuity of $V_\lambda$ with respect to $k_{\text{info}}$

**Lemma 2.18** (Monotonicity of $V_\lambda$ with respect to $k_{\text{info}}$)**.** *Let $V_\lambda$ be the value function of a given control problem with information cost parameter $k_{\text{info}}$. Changing the information cost parameter to $\tilde{k}_{\text{info}}$ with $\tilde{k}_{\text{info}} < k_{\text{info}}$ results in a value function $\tilde{V}_\lambda$ with*

$$\tilde{V}_\lambda(x) \leq V_\lambda(x) \quad \forall x \in \mathcal{S}.$$

*Proof.* Let $u^*$ be the optimal policy with respect to the parameter $k_{\text{info}}$, i.e. it holds $V_\lambda(x) = J_\lambda(x, u^*)$ for all $x$. For a fixed policy, the cost functional $J_\lambda$, defined in (2.3), is obviously monotone in $k_{\text{info}}$. This means that

$$\tilde{J}_\lambda(x, u^*) \leq J_\lambda(x, u^*),$$

where $\tilde{J}_\lambda$ is the cost functional given the parameter $\tilde{k}_{\text{info}}$. From $\tilde{V}_\lambda(x) \leq \tilde{J}_\lambda(x, u^*)$ $\forall x \in \mathcal{S}$ it directly follows $\tilde{V}_\lambda \leq V_\lambda$. □

**Theorem 2.19** (Continuity of $V_\lambda$ with respect to $k_{\text{info}}$ at $k_{\text{info}} > 0$)**.** *For each state $x \in \mathcal{S}$, the function $V_\lambda(x)$ is continuous in $k_{\text{info}} > 0$.*

*Proof.* We show continuity from right and left separately.
Continuity from the right: For a given $k_{\text{info}} > 0$ consider the corresponding optimal policy $u^*$ and the value function $V_\lambda$. Applying the policy $u^*$ for some $\tilde{k}_{\text{info}} = k_{\text{info}} + \delta > k_{\text{info}}$ , $\delta > 0$, increases only the information costs (as a part of $V_\lambda$), namely by the factor $\frac{\tilde{k}_{\text{info}}}{k_{\text{info}}} = \frac{k_{\text{info}}+\delta}{k_{\text{info}}} > 1$, such that for the corresponding cost functional it holds

$$\tilde{J}_\lambda(x, u^*) \leq \frac{k_{\text{info}} + \delta}{k_{\text{info}}} V_\lambda(x)$$

for any $x \in \mathcal{S}$. Let $\tilde{V}_\lambda$ denote the value function given the parameter $\tilde{k}_{\text{info}}$. From $\tilde{V}_\lambda(x) \leq \tilde{J}_\lambda(x, u^*)$ for all $x \in \mathcal{S}$ it follows

$$\tilde{V}_\lambda(x) \leq \frac{k_{\text{info}} + \delta}{k_{\text{info}}} V_\lambda(x) \quad \forall x \in \mathcal{S},$$

and, using the monotonicity of $V_\lambda$ with respect to $k_{\text{info}}$,

$$0 < \tilde{V}_\lambda(x) - V_\lambda(x) \leq \frac{\delta}{k_{\text{info}}} V_\lambda(x) \quad \forall x \in \mathcal{S}.$$

Hence, given an $\varepsilon > 0$ and $x \in \mathcal{S}$, we can choose $\delta < \frac{\varepsilon \cdot k_{\text{info}}}{V_\lambda(x)}$ to guarantee

$$|\tilde{V}_\lambda(x) - V_\lambda(x)| < \varepsilon.$$

Continuity from the left: Starting with a $k_{\text{info}} > 0$ we now consider $\tilde{k}_{\text{info}} = k_{\text{info}} - \delta < k_{\text{info}}$, $\delta > 0$, and the corresponding optimal policy $\tilde{u}$ as well as the value function $\tilde{V}_\lambda(x) = \tilde{J}_\lambda(x, \tilde{u})$. By the same arguments as before we get

$$V_\lambda(x) \leq \frac{k_{\text{info}}}{k_{\text{info}} - \delta} \tilde{V}_\lambda(x) \quad \forall x \in \mathcal{S},$$

and

$$0 < V_\lambda(x) - \tilde{V}_\lambda(x) \leq \frac{\delta}{k_{\text{info}}} V_\lambda(x).$$

Given $\varepsilon > 0$ and $x \in \mathcal{S}$, we choose again $\delta < \frac{\varepsilon \cdot k_{\text{info}}}{V_\lambda(x)}$ and get

$$|V_\lambda(x) - \tilde{V}_\lambda(x)| < \varepsilon$$

which completes the proof. □

**Theorem 2.20** (Continuity of $V_\lambda$ with respect to $k_{\text{info}}$ at $k_{\text{info}} = 0$)**.** *Given a Markov control problem with information costs, let $V_{\lambda,k_{\text{info}}} = V_{\lambda,k_{\text{info}}}(x)$ denote the value function depending on the parameter $k_{\text{info}}$. Let $V_{\lambda,0} = V_{\lambda,0}(x)$ denote the value function of the corresponding original Markov control problem (without information costs). It holds*

$$V_{\lambda,k_{\text{info}}}(x) \overset{k_{\text{info}} \to 0}{\longrightarrow} V_{\lambda,0}(x) \quad \forall x \in \mathcal{S}.$$

*Proof.* By Corollary 1.6 we can state that the value function $V_{\lambda,0}$ of the original control problem fulfills

$$V_{\lambda,0} = (\lambda I - L_{a_0})^{-1} c_{a_0},$$

where $a_0 : \mathcal{S} \to \mathcal{A}$ is the corresponding optimal policy and $L_{a_0}(x,y) := L_{a_0(x)}(x,y)$ as well as $c_{a_0}(x) := c(x, a_0(x))$ for all $x, y \in \mathcal{S}$. We will proceed as follows: Given for each state $x \in \mathcal{S}$ the optimal action $a_0(x) \in \mathcal{A}$ of the original control problem and a cost parameter $k_{\text{info}} > 0$, we define a policy $u : \mathcal{S} \to \mathcal{A} \times (0, \infty]$ for the information cost problem and calculate the corresponding cost functional $J_{\lambda,k_{\text{info}}}(x, u)$. We show that these cost functionals $J_{\lambda,k_{\text{info}}}$ converge to $V_{\lambda,0}$ for $k_{\text{info}} \to 0$. Then, we use the fact that

1. $V_{\lambda,k_{\text{info}}}(x) \leq J_{\lambda,k_{\text{info}}}(x,u)$ ($V_{\lambda,k_{\text{info}}}$ is optimal),

2. $V_{\lambda,k_{\text{info}}}(x) \geq V_{\lambda,0}(x)$ for all $k_{\text{info}} > 0$, $x \in \mathcal{S}$ ($V_{\lambda,0}$ is lower bound, see Section 2.2.2),

to see that $V_{\lambda,k_{\text{info}}}$ converges to $V_{\lambda,0}$ for $k_{\text{info}} \to 0$ as well.
Given $k_{\text{info}} > 0$, we define $\tau_0 := \sqrt{k_{\text{info}}}$ (independent of $x$) and consider the policy $u(x) = (a_0(x), \tau_0)$ which is in general not optimal. The corresponding cost functional is given by

$$J_{\lambda,k_{\text{info}}} = (I - D_{\tau_0}P_u)^{-1}(C_u + k_{\text{info}}e_{\tau_0}),$$

see Lemma 2.6, where $C_u(x) := C(x, a_0(x), \tau_0)$ (as in Lemma 2.17) and $P_u(x,y) = e^{L_{a_0(x)}\tau_0}$ for all $x, y \in \mathcal{S}$. Calculating the limit of this functional for $k_{\text{info}} \to 0$ leads to the following problem: with $k_{\text{info}} \to 0$ we get $\tau_0 \to 0$, and with it the matrix $I - D_{\tau_0}P_u$ converges to zero and is not invertible anymore. At the same time, the expression $C_u + k_{\text{info}}e_{\tau_0}$ converges to zero as well. The idea is to multiply the whole expression by the term $\frac{\tau_0}{\tau_0}$ which leads to

$$\begin{aligned} J_{\lambda,k_{\text{info}}} &= \frac{\tau_0}{\tau_0}(I - D_{\tau_0}P_u)^{-1}(C_u + k_{\text{info}}e_{\tau_0}) \\ &= \left(\frac{1}{\tau_0}(I - D_{\tau_0}P_u)\right)^{-1} \frac{1}{\tau_0}(C_u + k_{\text{info}}e_{\tau_0}). \end{aligned}$$

Now we analyze the matrix $\frac{1}{\tau_0}(I - D_{\tau_0}P_u)$. For a special row, referring to some state $x$, one can write

$$\begin{aligned} [D_{\tau_0}P_u](x,\cdot) &= \left[e^{(L_{a_0(x)}-\lambda I)\tau_0}\right](x,\cdot) \\ &= \left[\sum_{j=0}^{\infty} \frac{(L_{a_0(x)} - \lambda I)^j (\tau_0)^j}{j!}\right](x,\cdot) \end{aligned}$$

and

$$\begin{aligned} &\left[\frac{1}{\tau_0}(I - D_{\tau_0}P_u)\right](x,\cdot) \\ = &\left[\frac{1}{\tau_0}\left(I - \left(I + (L_{a_0(x)} - \lambda I)\tau_0 + \frac{1}{2}(L_{a_0(x)} - \lambda I)^2(\tau_0)^2 + \ldots\right)\right)\right](x,\cdot) \\ = &\left[-(L_{a_0(x)} - \lambda I) - \frac{1}{2}(L_{a_0(x)} - \lambda I)^2\tau_0 + \ldots\right](x,\cdot) \\ \overset{\tau_0\to 0}{\longrightarrow} &\ [\lambda I - L_{a_0 x)}](x,\cdot) \end{aligned}$$

For the whole matrix $\frac{1}{\tau_0}(I - D_{\tau_0}P_u)$ this means

$$\frac{1}{\tau_0}(I - D_{\tau_0}P_u) \quad \overset{\tau_0\to 0}{\longrightarrow} \quad \lambda I - L_{a_0}.$$

Next we analyze the term $\frac{1}{\tau_0}C_u$. We write $c_a(x) := c(x,a)$ for fixed $a \in \mathcal{A}$ and calculate

$$\begin{aligned} C(x,a,\tau) &= \mathbb{E}_x^a \left( \int_0^\tau e^{-\lambda s} c(X_s,a)\, ds \right) \\ &= \int_0^\tau e^{-\lambda s} \left[ e^{L_a s} c_a \right](x)\, ds \\ &= \left[ (L_a - \lambda I)^{-1} \left( e^{(L_a - \lambda I)\tau} - I \right) c_a \right](x). \end{aligned}$$

This implies

$$\begin{aligned} \frac{1}{\tau_0} C_u(x) &= \frac{1}{\tau_0} \left[ (L_{a_0} - \lambda I)^{-1} \left( e^{(L_{a_0} - \lambda I)\tau_0} - I \right) c_{a_0} \right](x) \\ &= \frac{1}{\tau_0} \left[ (L_{a_0} - \lambda I)^{-1} \left( (L_{a_0} - \lambda I)\tau_0 + \frac{1}{2}(L_{a_0} - \lambda I)^2 (\tau_0)^2 + ... \right) c_{a_0} \right](x) \\ &= \frac{1}{\tau_0} \left[ \left( \tau_0 + \frac{1}{2}(L_{a_0} - \lambda I)(\tau_0)^2 + ... \right) c_{a_0} \right](x) \\ &= \left[ \left( 1 + \frac{1}{2}(L_{a_0} - \lambda I)\tau_0 + ... \right) c_{a_0} \right](x) \\ &\overset{\tau_0 \to 0}{\longrightarrow} c_{a_0}(x), \end{aligned}$$

such that, for the whole vector:

$$\frac{1}{\tau_0} C_u \overset{\tau_0 \to 0}{\longrightarrow} c_{a_0}.$$

In a final step, we consider the expression $\frac{1}{\tau_0} k_{\text{info}} e_{\tau_0}$, replacing $k_{\text{info}} = (\tau_0)^2$:

$$\begin{aligned} \frac{1}{\tau_0} k_{\text{info}} e_{\tau_0} &= \frac{1}{\tau_0} (\tau_0)^2 e^{-\lambda \tau_0} \\ &= \tau_0 e^{-\lambda \tau_0} \\ &\overset{\tau_0 \to 0}{\longrightarrow} 0. \end{aligned}$$

We can put everything together and get (as $k_{\text{info}} \to 0$ is equivalent to $\tau_0 \to 0$):

$$J_{\lambda, k_{\text{info}}} \overset{k_{\text{info}} \to 0}{\longrightarrow} (\lambda I - L_{a_0})^{-1} c_{a_0} = V_{\lambda,0}.$$

Now, from

$$V_{\lambda,0} \le V_{\lambda,k_{\text{info}}} \le J_{\lambda,k_{\text{info}}} \overset{k_{\text{info}} \to 0}{\longrightarrow} V_{\lambda,0},$$

the convergence of $V_{\lambda,k_{\text{info}}}$ immediately follows. □

Figure 2.8 shows the value function for the 2-state-example depending on the information cost parameter $k_{\text{info}}$.

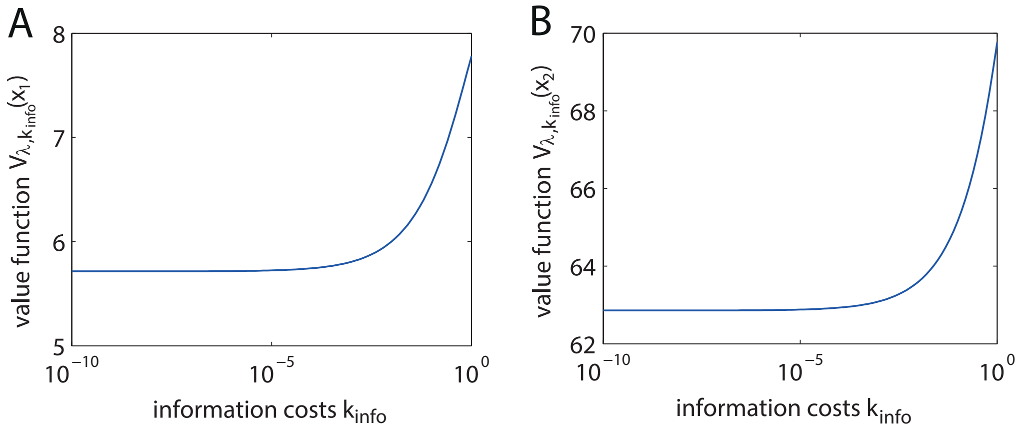

Figure 2.8: **$k_{\text{info}}$ vs. optimal discounted costs for the 2-state-example.** The value function $V_{\lambda,k_{\text{info}}}$ for Example 2.11 depending on the information costs $k_{\text{info}}$ in a logarithmic scale, evaluated at state $x_1$ (panel A) and state $x_2$ (panel B). For $k_{\text{info}} = 1$ the values agree with those given in the first row of Table 2.1 ($V_{\lambda,k_{\text{info}}}(x_1) = 7.78$, $V_{\lambda,k_{\text{info}}}(x_2) = 69.77$). For $k_{\text{info}} = 10^{-10}$ the values are close to the values $V_{\lambda,0}(x_1) = 5.71$ and $V_{\lambda,0}(x_2) = 62.86$ of the original Markov control model without information costs.

Now we turn to the last two questions that were formulated in the beginning of this section: How does the optimal lag time $\tau^*$ depend on the information costs $k_{\text{info}}$?

### Connection between $k_{\text{info}}$ and $\tau^*$

Knowing that the value function is monotone and continuous in $k_{\text{info}}$, we now analyze the structure of the optimal policy depending on $k_{\text{info}}$. How do the optimal lag times $\tau^*(x)$ change when $k_{\text{info}}$ changes? Intuitively, a reduction of the information costs should lead to a higher frequency of tests, or, equivalently, to smaller lag times. However, as the following simple example shows, such a monotonicity does not hold in general.

**Example 2.21** (No monotonicity in $\tau^*$). *We consider again a 3-state-model similar to the one described in Example 2.12. We choose*

$$L_1 = \begin{pmatrix} -0.1 & 0.1 & 0 \\ 0.1 & -0.2 & 0.1 \\ 0 & 0 & 0 \end{pmatrix}, \quad L_2 = \begin{pmatrix} -0.1 & 0.1 & 0 \\ 1 & -1.1 & 0.1 \\ 0 & 15 & -15 \end{pmatrix},$$

*as well as $c_{\mathcal{S}}(x_1) = c_{\mathcal{S}}(x_I) = 0$, $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_1) = 0$, $c_{\mathcal{A}}(a_2) = 2$ and $\lambda = 0.1$. Again, $x_2$ is the "bad" state producing a lot of state costs, and $a_2$ is the expensive action driving the process quickly out of this "bad" state and towards the "safe" state $x_1$, while for the free action $a_1$, state $x_2$ is absorbing.*

*We calculate the optimal policy for different $k_{\text{info}}$ and observe the following structure. It holds $a^*(x_1) = a_1$ and $a^*(x_2) = a_2$ for all $k_{\text{info}} > 0$, whereas for the intermediate*

*state the optimal action depends on $k_{\text{info}}$. For $k_{\text{info}} < 0.48$ the optimal action is given by $a^*(x_I) = a_1$, while for $k_{\text{info}} \geq 0.48$ it holds $a^*(x_I) = a_2$. For $x_1$ and $x_2$ the optimal lag time $\tau^*$ is increasing in $k_{\text{info}}$. However, for the intermediate state $x_I$ there is a point of discontinuity at $k_{\text{info}} = 0.48$: With the switchover in the optimal action, the optimal lag time decreases in a volatile way. Only for areas of constant action the lag time $\tau^*(x_I)$ increases with $k_{\text{info}}$. At $k_{\text{info}} = 6.55$ there is another point of discontinuity in $\tau^*$ for $x_I$ and $x_2$: From $\tau^*(x_I) \approx \tau^*(x_2) \approx 9.3$ it jumps to $\tau^* = \infty$ for all $k_{\text{info}} > 6.55$, see Figure 2.9.*
*The corresponding value functions are shown in Figure 2.10.*

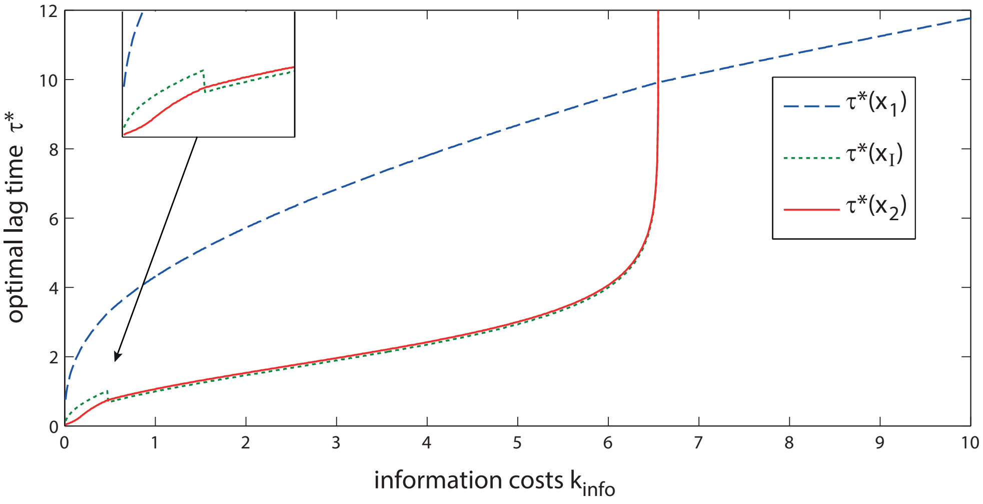

Figure 2.9: **$k_{\text{info}}$ vs. optimal lag times for the 3-state-example.** The optimal lag times $\tau^*(x)$ for Example 2.21 depending on the information costs $k_{\text{info}}$. At $k_{\text{info}} = 0.48$ the optimal lag time $\tau^*(x_I)$ of the intermediate state $x_I$ performs a jump.

*Interpretation: As long as tests are not too expensive, the optimal policy consists of choosing the free action $a_1$ for state $x_I$. Then, a test will follow after a short time interval of unobserved progress, such that in case of a transition to $x_2$ the action can be adapted without too much loss in time. This way, the process is prevented from staying for a longer time in state $x_2$. In case of larger $k_{\text{info}}$ this kind of safeguarding is not affordable such that a better choice (starting from $x_I$) is to choose action $a_2$, which leads the process back to the "safe" state $x_1$ with a certain (high) probability. In order to avoid too many action costs, a (more or less) quickly following test indicates whether the process returned to state $x_1$ such that the action can be adapted.*

*Analysis of the points of discontinuity: In order to understand what is happening at the breaking points $k_{\text{info}} = 0.48$ and $k_{\text{info}} = 6.55$, we calculate some "conditional" optimal policies, i.e. policies which are optimal within a set of policies fulfilling special conditions.*

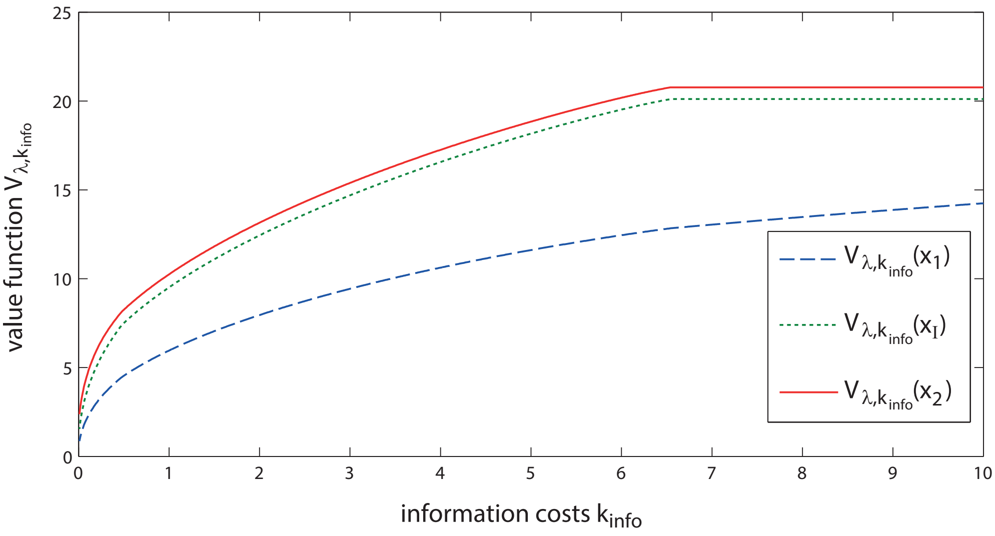

Figure 2.10: **$k_{\text{info}}$ vs. optimal discounted costs for the 3-state-example.** The value function $V_{\lambda,k_{\text{info}}}$ for Example 2.21 depending on the information costs $k_{\text{info}}$.

*The conditions are*

*A: The action chosen for $x_I$ has to be $a_1$ and the lag time has to be finite, i.e. $a(x_I) = a_1, \tau(x_I) < \infty$.*

*B: The action chosen for $x_I$ has to be $a_2$ and the lag time has to be finite, i.e. $a(x_I) = a_2, \tau(x_I) < \infty$.*

*C: The lag time chosen for $x_I$ has to be $\infty$, i.e. $\tau(x_I) = \infty$.*

*The sets of policies referring to these three conditions form a disjoint and full partition of the overall set of policies.*
*The corresponding conditional value functions for $x_I$ are shown in Figure 2.11. There are two points of intersection where the minimum property switches from one of these functions to another. These intersection points coincide with the given breaking points $k_{\text{info}} = 0.48$ and $k_{\text{info}} = 6.55$. The connection should be clear: The overall optimal policy is the minimum of the conditional optimal policies, and the overall value function $V_\lambda$ is the minimum of the conditional value functions.*

By means of Example 2.21 we see that the optimal lag time $\tau^*$ is in general not monotone in $k_{\text{info}}$. At the same time, this example refutes the continuity of $\tau^*$ with respect to $k_{\text{info}}$, compare the breaking points $k_{\text{info}} = 0.48$ and $k_{\text{info}} = 6.55$ where its value performs a jump. Still it seams that for $k_{\text{info}} \to 0$ the optimal lag time converges to 0 which would again, as in the considerations for the value function (Theorem 2.20), give a reasonable connection to the original Markov control process (without information costs). However, a simple argument shows that such

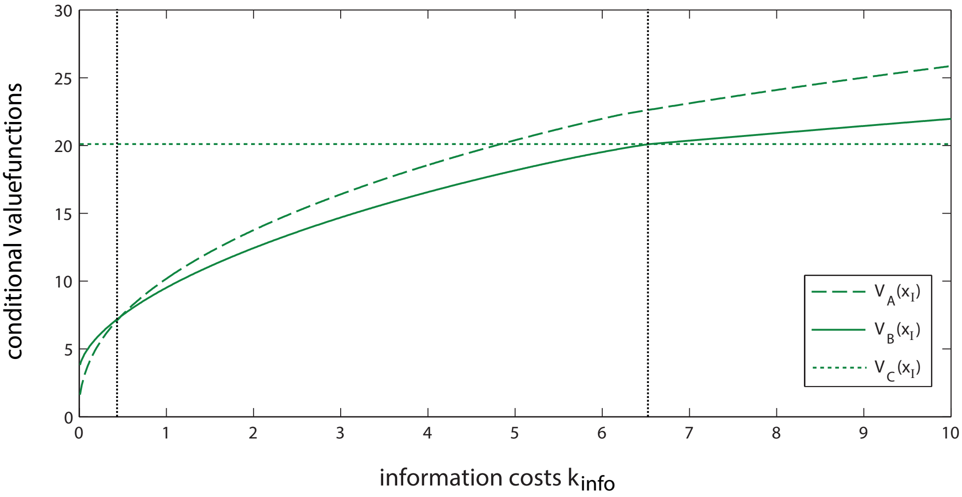

Figure 2.11: **$k_{\text{info}}$ vs. conditional value functions.** Conditional value functions $V_A$, $V_B$ and $V_C$, evaluated at the intermediate state $x_I$, for the 3-state-example 2.21 and the conditions described therein. The vertical lines indicate the intersection points at which the minimum property switches from one conditional value function to another. These points coincide with the breaking points $k_{\text{info}} = 0.48$ and $k_{\text{info}} = 6.55$.

a continuity at $k_{\text{info}} = 0$ is not given in general: Let us assume that there exists a state $x \in \mathcal{S}$ which is absorbing for all available actions, i.e. once the process enters this state it will never leave it again, no matter which action is chosen. Independent of the given $k_{\text{info}} > 0$, the optimal lag time for this state will be $\tau^*(x) = \infty$, as no further observation is reasonable. Excluding this situation, we can indeed formulate

**Theorem 2.22** (Continuity of $\tau^*$ with respect to $k_{\text{info}}$ at $k_{\text{info}} = 0$)**.** *Assume the action space $\mathcal{A}$ to be finite. Let $a_0$ denote the optimal policy (i.e. optimal choice of action) of the original control problem (without information costs). For $x \in \mathcal{S}$ assume that there exists a state $y \in \mathcal{S}$ with $a_0(x) \neq a_0(y)$ and $\mathbb{P}_{a_0(x)}(X_t = y|X_0 = x) > 0$ for some $t > 0$.*
*Then it holds*

$$\tau^*_{k_{\text{info}}}(x) \overset{k_{\text{info}}\to 0}{\longrightarrow} 0,$$

*where $\tau^*_{k_{\text{info}}}(x)$ is the optimal lag time for state $x$ given the information costs $k_{\text{info}}$.*

*Proof.* We seek a proof per contradiction.
Assuming that $\tau^*_{k_{\text{info}}}(x)$ does not converge to zero for some $x \in \mathcal{S}$ means

$$\exists \varepsilon > 0 \;\; \forall \delta > 0 \;\; \exists k_{\text{info}} < \delta : \;\; \tau^*_{k_{\text{info}}}(x) \geq \varepsilon.$$

In other words: For this $x$, there exist an $\varepsilon > 0$ and a sequence $(k^n_{\text{info}})_{n\in\mathbb{N}}$ with $k^n_{\text{info}} \to 0$ but $\tau^*_{k^n_{\text{info}}}(x) \geq \varepsilon \; \forall n$. Use the short notation $\tau^*_n := \tau^*_{k^n_{\text{info}}}$. Let $V_{\lambda,n}$ be the value function referring to $k^n_{\text{info}}$, and let $a^*_n$ resp. $L^*_n$ be the corresponding optimal

choice of action resp. the suitable generator. We know from Theorem 2.20 that $V_{\lambda,n} \overset{n\to\infty}{\longrightarrow} V_{\lambda,0}$, where $V_{\lambda,0}$ is the value function of the original control problem.
We now consider the time interval $[0,\varepsilon)$ and analyze for each $n$ the difference between the control $u_n^* = (a_n^*, \tau_n^*)$ and $a_0$. For this purpose, we imagine switching at time $\varepsilon$ from the policy $u_n^* = (a_n^*, \tau_n^*)$ to the policy $a_0$ (assuming that from time $\varepsilon$ on information about the process is free of charge), see Figure 2.12, which would result in costs $w_\varepsilon(x, a_n^*(x))$, where

$$w_\varepsilon(x,a) := \mathbb{E}_x^a \left( \int_0^\varepsilon e^{-\lambda s} c(X_s, a)\, ds \right) + \left( e^{-\lambda\varepsilon} e^{L_a \varepsilon} V_{\lambda,0} \right)(x).$$

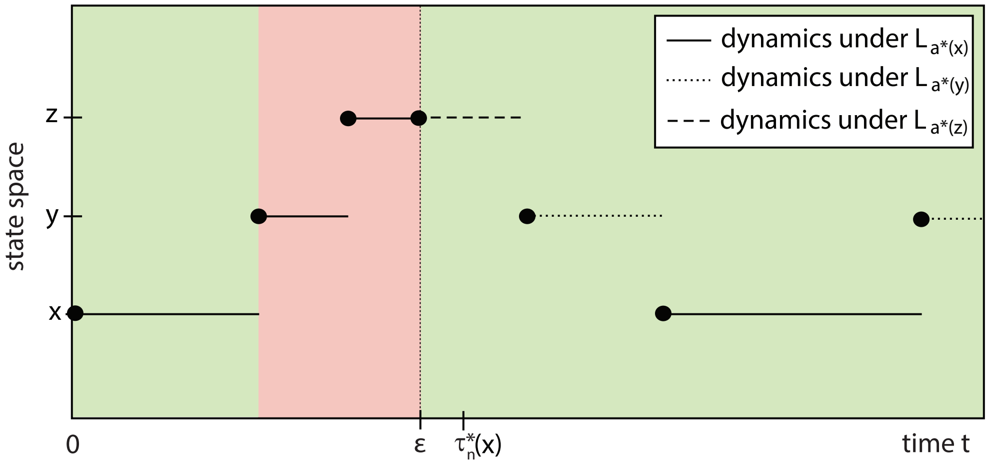
dynamics under $L_{a^*(x)}$
dynamics under $L_{a^*(y)}$
dynamics under $L_{a^*(z)}$
state space
z
y
x
0
ε
$\tau_n^*(x)$
time t

Figure 2.12: **Controlled Markov process.** Possible trajectory given that the process is controlled according to the policy $u_n^* = (a_n^*, \tau_n^*)$ before time $\varepsilon$ and according to $a_0$ after time $\varepsilon$ (assuming for simplicity that $a_n^*(x) = a_0(x)$). As in Figure 2.7, the green areas mark the periods of time where the process is optimally controlled according to the actual state, while the red area marks the period of time where the control is not adapted to the actual state. The dynamics of this switched system do not depend on $\tau_n^*$ as it holds $\tau_n^* > \varepsilon$.

Noting that such a switching to the (overall) optimal policy $a_0$ at time $\varepsilon$ can only lead to an improvement in the total costs makes clear that it holds

$$V_{\lambda,n}(x) \geq w_\varepsilon(x, a_n^*(x)).$$

At the same time, we have

$$w_\varepsilon(x,a) > V_{\lambda,0}(x)$$

for all $a \in \mathcal{A}$. In order to see this we make the following distinction of cases.
<u>Case 1:</u> $a = a_0(x)$. In this case, starting from state $x$ at time $t = 0$, the dynamics of the processes conducted by $a$ and $a_0$, respectively, coincide as long as there is no jump to a state $y \in \mathcal{S}$ with $a_0(y) \neq a_0(x)$. However, given the existence of such a $y$ with $\mathbb{P}_{a_0(x)}(X_t = y | X_0 = x) > 0$ for *some* $t > 0$ (see assumptions) it follows

$\mathbb{P}_{a_0(x)}(X_t = y|X_0 = x) > 0$ for *all* $t > 0$. (This is due to the fact that the jumping times are exponentially distributed and can be arbitrarily small or large, while the transition probabilities do not depend on these jumping times.) Especially, it holds $\mathbb{P}_{a_0(x)}(X_t = y \text{ for some } t \in (0,\varepsilon)|X_0 = x) > 0$, which means that with a positive probability the dynamics of the two processes during the time interval $[0,\varepsilon)$ are different.

Case 2: $a \neq a_0(x)$. In this case, from time zero on the difference in the action leads to a difference in the dynamics.

For both cases, the difference in the dynamics during the time interval $[0,\varepsilon)$ together with the optimality of the control $a_0$ yields

$$\begin{aligned} w_\varepsilon(x,a) &> \mathbb{E}_x^{a_0}\left(\int_0^\varepsilon e^{-\lambda s} c(X_s, a_0(X_s))\, ds\right) + \left(e^{-\lambda\varepsilon} e^{L_{a_0}\varepsilon}\right) V_{\lambda,0}(x) \\ &= V_{\lambda,0}(x). \end{aligned}$$

Now we use the fact that $\mathcal{A}$ is finite to define (for fixed $x \in \mathcal{S}$)

$$d := \min_{a\in\mathcal{A}} \left\{w_\varepsilon(x,a) - V_{\lambda,0}(x)\right\}.$$

By

$$V_{\lambda,n}(x) \geq w_\varepsilon(x, a_n^*(x)) > V_{\lambda,0}(x)$$

it follows

$$V_{\lambda,n}(x) - V_{\lambda,0}(x) \geq w_\varepsilon(x, a_n^*(x)) - V_{\lambda,0}(x) \geq d > 0$$

for all $n \in \mathbb{N}$, which is a contradiction to $V_{\lambda,n} \overset{n\to\infty}{\longrightarrow} V_{\lambda,0}$. □

**Remark 2.23** (Monotonicity of $\tau^*$ for fixed actions)**.** *In Example 2.21 the jump of the optimal lag time at the critical value $k_{\text{info}} = 0.48$ is connected to a switch in the optimal action $a^*(x_I)$. However, for those values of $k_{\text{info}}$ where the optimal choice of action is constant the optimal lag time seems indeed to be monotone. The following argumentation confirms this observation. Clearly, the information costs*

$$J_{\text{info}}(x,u) = \mathbb{E}_x^u\left(\sum_{\substack{j\in\mathbb{N}_0\\ t_j<\infty}} e^{-\lambda t_{j+1}} k_{\text{info}}\right)$$

*for a given policy $u$ decrease with increasing $\tau(y)$ for all states $y \in \mathcal{S}$, i.e. it holds*

$$J_{\text{info}}(x,\tilde{u}) \leq J_{\text{info}}(x,u) \quad \text{for } \tilde{\tau}(y) > \tau(y),$$

*where the policy $\tilde{u}$ coincides with $u$ for all parameters except the lag time $\tilde{\tau}(y) > \tau(y)$. Given an optimal policy $u^*$ with finite $\tau^*(x)$ for all $x \in \mathcal{S}$, we can conclude that the net costs $J_{\text{net}}(x,u^*) = J_\lambda(x,u^*) - J_{\text{info}}(x,u^*)$ have to be increasing with increasing $\tau(y)$; otherwise $\tau^*(y)$ could be increased in order to minimize the total costs $J_\lambda(x,u)$. Given a value $k_{\text{info}}$ of information costs, the optimal lag time in some sense defines the right "equilibrium" between decreasing information costs and increasing net costs.*

*Meanwhile, the parameter $k_{\text{info}}$ can be seen as a weighting factor identifying the impact of the information costs on this equilibrium. Hence, as long as we stick to the same choice of actions, a smaller value of $k_{\text{info}}$ reduces the impact of the information costs and causes smaller optimal lag times.*

### 2.2.4 Sensitivity with respect to Lag Times $\boldsymbol{\tau(x)}$

In the last section we analyzed how the value function and the optimal policy depend on the information cost parameter $k_{\text{info}}$. While the value function was discovered to be monotone and continuous in $k_{\text{info}}$, the optimal lag times can perform jumps which can go both downward and upward.
Another question of interest concerns the sensitivity of the value function with respect to deviations from the optimal lag time $\tau^*$: Given the optimal policy, how do small changes in the lag times affect the cost functional? In real-world applications it might be an evident problem that an exact compliance with the prescribed lag times is not possible.

In fact, the answer to this question was already given in Lemma 2.10: The continuity of the cost functional $J_\lambda^u$ with respect to the lag time $\tau$ implies that "small" changes in $\tau$ cannot have a crucial effect on the expected costs. Of course, what "small" means depends on the concrete example. In order to get an impression of how the cost functional behaves depending on $\tau$, we consider again the 2-state-example 2.11. The graphics in Figure 2.13 show the cost functionals $J_\lambda^u$ and $J_{\text{net}}^u = J_{\text{net}}(\cdot, u)$ for $u(x) = (a(x), \tau(x))$ as functions of $\tau(x_1)$ (resp. $\tau(x_2)$), while the other lag time $\tau(x_2)$ (resp. $\tau(x_1)$) is fixed. The actions are fixed to be optimal.

As for Figure 2.13 we can make the following observations. While the net costs $J_{\text{net}}^u$ (total cost minus information costs) are monotonously increasing in $\tau$ and do not attain a minimum (which agrees with Remark 2.23), the total costs $J_\lambda^u$ exhibit a unique minimum. The minimum points $\tau^*$ of $J_\lambda^u(x_1)$ and $J_\lambda^u(x_2)$ coincide. Panel A shows that the costs are nearly constant for a wide area of values of $\tau(x_1)$, roughly $3 \leq \tau(x_1) \leq 20$. The response towards changes in $\tau(x_2)$ is more sensitive, see panel B: The value of the cost functional $J_\lambda(x_2)$ clearly increases when $\tau(x_2)$ differs from its optimum $\tau^*(x_2) = 1.8$.
Choosing a non optimal combination of actions (e.g. $a(x_1) = 2$, $a(x_2) = 1$) for this example results in monotonously decreasing cost functionals without minimum. That is, in this case the optimal lag times would be infinite.

#### Numerical consequences

In the case of a low sensitivity of the cost functional with respect to the lag time parameter $\tau$ – as it is given in panel A of Figure 2.13 – the problem of finding the minimum solution is numerically ill-conditioned. A gradient descent with respect to $\tau$ would be extremely slow because the gradient would almost vanish within a wide area around the minimum solution. However, as mentioned on page 55, many applications suggest to discretize the domain of the parameter $\tau$ (because

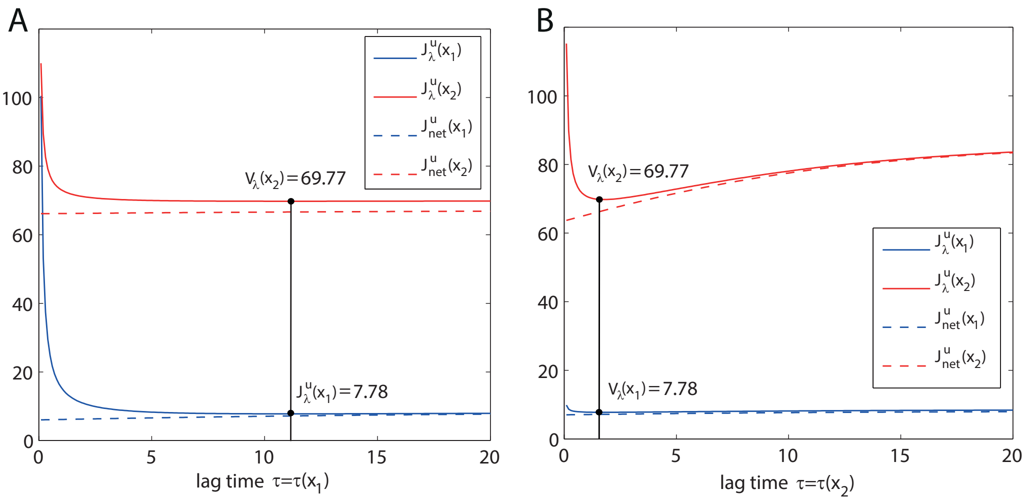

Figure 2.13: **Sensitivity with respect to $\tau$ for the 2-state-example 2.11.** A: Lag time $\tau(x_1)$ vs. cost functionals $J^u_\lambda$ and $J^u_{\text{net}}$ for fixed $\tau(x_2) = 1.8$ and $a(x_1) = 1$, $a(x_2) = 2$. B: Lag time $\tau(x_2)$ vs. cost functionals $J^u_\lambda$ and $J^u_{\text{net}}$ for fixed $\tau(x_1) = 11.3$ and $a(x_1) = 1$, $a(x_2) = 2$.
All other parameters coincide with those given in the first line of Table 2.1, i.e. $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_2) = 2$, $k_{\text{info}} = 1$, $\lambda = 0.1$. The minima of $J^u_\lambda(x_1)$ and $J^u_\lambda(x_2)$ are attained at $\tau^*(x_1) = 11.3$ (panel A) resp. $\tau^*(x_2) = 1.8$ (panel B) with $J^u_\lambda(x_1) = V_\lambda(x_1) = 7.78$ and $J^u_\lambda(x_2) = V_\lambda(x_2) = 69.77$ which is consistent with the values of the optimal policy declared in Table 2.1.

tests cannot be placed arbitrarily exact in time), which overcomes this difficulty. Furthermore, for practical purpose, a nearby solution is completely satisfying as long as the resulting costs are close to optimal.

## 2.3 The Average-Cost Criterion

So far we deduced an optimality equation for the discounted-cost criterion in the new setting of Markov control with information costs; we extended the analysis by calculating a cost splitting and compared the net-costs to the optimal costs of the original control problem. Moreover, we considered the information cost parameter $k_{\text{info}}$ as a variable and explored how the value function and the optimal lag times depend on this parameter. Finally, we studied the sensitivity of the expected discounted costs with respect to deviations from the optimal lag times.
All these steps will now be reproduced for the average-cost criterion. Again we will restrict the analysis to finite state spaces. Interestingly, the approach will be completely different. Instead of directly analyzing the new cost functional, we will construct a freely observable Markov decision process which – as for the expected average costs – is equivalent to the given Markov decision process with information costs. Due to this preparatory work, we can apply all the results of Section 1.1.2

(the average-cost criterion in the original setting) to deduce an optimality equation for the new setting without further effort. The subsequent calculation of a cost splitting will be straightforward, and the monotonicity and continuity analysis will conform to the case of discounted costs.

### 2.3.1 The Average-Cost Optimality Equation

Recall that for a Markov decision process without information costs the average-cost criterion is given by

$$\limsup_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T c(X_s, u(X_s))\, ds \right),$$

compare Section 1.1. For ergodic dynamics, this value does not depend on $x \in \mathcal{S}$ but is given by a constant $\eta_u = \langle \mu_u, c_u \rangle$, where $\mu_u$ is the unique equilibrium distribution of the process under control $u$. The constant $\eta_u$ fulfills the equation

$$\eta_u = c(x, u(x)) + (L_u v)(x) \quad \forall x \in \mathcal{S},$$

where $v$ is a real-valued function on $\mathcal{S}$ which has no direct interpretation but forms a weighting factor between the different states, see Lemma 1.13. In the following, we will deduce an equivalent equation for the case of information costs and the cost functional

$$\bar{J}(x,u) = \limsup_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<T}} \left( \int_{t_j}^{t_{j+1}\wedge T} c(X_s, a(X_{t_j}))\, ds + k_{\text{info}} \right) \right), \tag{2.24}$$

where $t_{j+1} \wedge T := \min\{t_{j+1}, T\}$ as well as $t_0 = 0$, $t_{j+1} = t_j + \tau(X_{t_j})$ for $j \in \mathbb{N}$ and $u(x) = (a(x), \tau(x))$, compare (2.5).

#### Ergodic dynamics and finite lag times

For the case of ergodic dynamics, we tend to express the cost functional $\bar{J}(x,u)$ for a given policy $u \in \mathcal{U}$ with **finite** lag times (i.e. $\tau(x) < \infty$ for all $x \in \mathcal{S}$) in terms of an equilibrium distribution of the process $(X_t)_{t\geq 0}$.[5] However, the controlled process $(X_t)_{t\geq 0}$ itself is not a Markov process: Which generator determines the dynamics of the process at time $t$ depends on the last observation $X_{t_n}$ with $t_n = \max\{t_j : j \in \mathbb{N}, t_j \leq t\}$ and not on the actual state $X_t$. In other words, the past (and not only the present) is relevant for the future evolution of the process. It is therefore not clear how an equilibrium distribution could be characterized.
Being aware that also the costs produced by the process at time $t$ depend on the

[5] For infinite lag times the long-term average costs depend on the initial state and cannot be expressed in terms of an equilibrium distribution of the process $(X_t)_{t\geq 0}$.

last observation $X_{t_n}$, the idea is to consider the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ which is itself a Markov process with discrete index. Its transition matrix $P_u$ is given by

$$P_u(x,y) = e^{L_{a(x)}\tau(x)}(x,y) \quad \text{for all } x,y \in \mathcal{S}. \tag{2.25}$$

Let us denote the equilibrium distribution with respect to $P_u$ by $\mu$, i.e. we assume $\mu \in \mathbb{R}^{|\mathcal{S}|}$ to be a probability vector with

$$\mu P_u = \mu.$$

Each time the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ is in state $x \in \mathcal{S}$, the costs per unit of time during the following time interval $[t_j, t_j + \tau(x))$ of constant control are given by

$$C(x, a(x), \tau(x)) := \mathbb{E}_x^{a(x)} \left( \frac{1}{\tau(x)} \int_0^{\tau(x)} c(X_s, a(x))\, ds \right). \tag{2.26}$$

In order to include the information costs $k_{\text{info}}$ appearing at the end of this time interval, we define

$$\tilde{C}(x, a(x), \tau(x)) := C(x, a(x), \tau(x)) + \frac{k_{\text{info}}}{\tau(x)}. \tag{2.27}$$

In other words, this is the cost rate for all times $t \geq 0$ at which the last observation of the underlying process $(X_t)_{t\geq 0}$ has been $x$. Now the question is: What is the average proportion of time for this situation to appear? Naturally, we can calculate this proportion – let us denote it by $\tilde{\mu}(x)$ – by multiplying the value of the equilibrium distribution $\mu(x)$ (specifying how *often* this situation appears) by the lag time $\tau(x)$ (denoting how *long* the situation remains), followed by a scaling with respect to the weighted average of lag times, i.e. it holds

$$\tilde{\mu}(x) = \frac{\mu(x)\tau(x)}{\sum_{y\in\mathcal{S}} \mu(y)\tau(y)}. \tag{2.28}$$

In order to verify this equation, we can find an analytic expression for $\tilde{\mu}$, namely

$$\tilde{\mu}(x) = \lim_{n\to\infty} \mathbb{E}_y^u \left( \frac{1}{t_n} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{1}_{\{X_{t_j}=x\}}\, ds \right),$$

which – for ergodic dynamics – is independent of the initial state $y \in \mathcal{S}$. As it holds $\frac{1}{t_n}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}} \mathbb{1}_{\{X_{t_j}=x\}}\, ds \leq 1$ for all $n \in \mathbb{N}$, by the dominated convergence theorem we can take the limit into the expectation value and write

$$\tilde{\mu}(x) = \mathbb{E}_y^u \left( \lim_{n\to\infty} \frac{1}{t_n} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{1}_{\{X_{t_j}=x\}}\, ds \right).$$

Noting that

$$\int_{t_j}^{t_{j+1}} \mathbb{1}_{\{X_{t_j}=x\}}\, ds = \begin{cases} \tau(x) & \text{if } X_{t_j} = x, \\ 0 & \text{otherwise ,} \end{cases}$$

we can state that, by the strong law of large numbers,

$$\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}} \mathbb{1}_{\{X_{t_j}=x\}}\, ds \;\sim\; n\cdot\mu(x)\tau(x) \quad (a.s.) \quad \text{for } n\to\infty.$$

At the same time, we have

$$t_n \;\sim\; n\cdot\sum_{y\in\mathcal{S}}\mu(y)\tau(y) \quad (a.s.) \quad \text{for } n\to\infty,$$

and putting both together we get (2.28).
We can now calculate the long-term average costs for the given policy $u$ by taking the $\tilde{\mu}$-weighted average of $\tilde{C}$:

$$\bar{J}(x,u) = \sum_{y\in\mathcal{S}}\tilde{\mu}(y)\tilde{C}(y,u(y)) = \langle\tilde{\mu},\tilde{C}_u\rangle =: \eta_u \quad \text{for all } x\in\mathcal{S} \tag{2.29}$$

with $\tilde{C}_u(x) := \tilde{C}(x,u(x))$.

### An equivalent freely observable Markov decision process

Consider the Markov control model with information costs (2.1) and let $(X_t)_{t\geq 0}$ be the controlled process given a policy $u\in\mathcal{U}$. The idea is to formulate another control process $(Y_t)_{t\geq 0}$ which is freely observable but has the same long-term average costs as the process $(X_t)_{t\geq 0}$.

To this end, we consider the process $(t_j, X_{t_j})_{j\in\mathbb{N}_0}$ of observation times and observations of the given process $(X_t)_{t\geq 0}$. In a first step, we again consider finite lag times. What is the expected time the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ stays in some state $x\in\mathcal{S}$ before switching to another state $y\neq x$ when action $a\in\mathcal{A}$ and lag time $\tau\in(0,\infty)$ are chosen? That is, what is the expectation value of the "residence time"

$$r(x) := \min\{t_j : j\in\mathbb{N}, X_{t_j}\neq x\}$$

given that $X_0 = x$? As the underlying process $(X_t)_{t\geq 0}$ can still or again be in state $x$ after time $\tau$, this residence time can be any multiple of $\tau$. The number of time intervals of length $\tau$ that pass before the state of the observation process changes for the first time after starting in $x\in\mathcal{S}$ is geometrically distributed with parameter $p(x) := 1 - e^{L_a\tau}(x,x)$, and so it holds

$$\mathbb{E}(r(x)) = \frac{\tau}{p(x)} = \frac{\tau}{1-e^{L_a\tau}(x,x)}.$$

Under the condition that a transition takes place, the transition probabilities for the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ are given by

$$\frac{e^{L_a\tau}(x,y)}{\sum_{\tilde{y}\in\mathcal{S},\tilde{y}\neq x} e^{L_a\tau}(x,\tilde{y})}.$$

Adopting these characteristics of the observation process, we define for each action $a \in \mathcal{A}$ and each lag time $\tau \in (0, \infty)$ a generator $G_{a,\tau} \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ by

$$G_{a,\tau}(x,y) := \frac{1}{\tau} e^{L_a \tau}(x,y) \quad \text{for } y \neq x, \quad G_{a,\tau}(x,x) := -\frac{1}{\tau}(1 - e^{L_a \tau}(x,x)). \tag{2.30}$$

For $\tau = \infty$ we set

$$G_{a,\infty}(x,y) := 0 \quad \forall y \in \mathcal{S} \tag{2.31}$$

which is convenient in the sense that the observation process $(X_{t_j})_{j \in \mathbb{N}_0}$ will never leave a state $x \in \mathcal{S}$ with $\tau(x) = \infty$ because no further tests are made. (The underlying process $(X_t)_{t \geq 0}$ of course switches between the states as usual.) Moreover, by this choice we obtain a continuity

$$\lim_{\tau \to \infty} G_{a,\tau} = G_{a,\infty} \tag{2.32}$$

which will be relevant for future statements.

Now, we consider the process $(Y_t)_{t \geq 0}$ to be a completely observable Markov decision process with state space $\mathcal{S}$, action space $\mathcal{A} \times (0, \infty]$ and set of generators $\{G_{a,\tau} : a \in \mathcal{A}, \tau \in (0, \infty]\}$. Given $a \in \mathcal{A}$ and $\tau \in (0, \infty]$, the dynamics of the process $(Y_t)_{t \geq 0}$ are determined by $G_{a,\tau}$. That is, the control parameters stay the same, however, their interpretation changes: $\tau$ has no longer an interpretation of a lag time between observations but only – together with $a$ – determines the generator. The process $(Y_t)_{t \geq 0}$ is freely observable at all times and the generator is adapted as soon as a transition takes place, i.e. for a given policy $u(x) = (a(x), \tau(x))$ the process is driven by the generator $G_u$ with

$$G_u(x,y) := G_{a(x),\tau(x)}(x,y) \quad \text{for all } x, y \in \mathcal{S}. \tag{2.33}$$

In terms of the transition matrix $P_u$ defined in (2.25) it holds, as for finite $\tau(x)$,

$$G_u = \begin{pmatrix} \ddots & & 0 \\ & \frac{1}{\tau(x)} & \\ 0 & & \ddots \end{pmatrix} (P_u - I), \tag{2.34}$$

where $I \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ is the identity matrix. By interpreting $\frac{1}{\infty} := 0$ this equation holds for infinite lag times, as well, no matter how the corresponding entries of the transition matrix $P_u$ are defined.

The two processes $(Y_t)_{t \geq 0}$ and $(X_t)_{t \geq 0}$ are completely independent of each other. However, as for the average dynamics, the process $(Y_t)_{t \geq 0}$ can be seen as the continuous analogue of the observation process $(X_{t_j})_{j \in \mathbb{N}_0}$: By construction, the expected residence times coincide for these two processes, and a transition of the process $(Y_t)_{t \geq 0}$ to another state refers to getting a new information about the process $(X_t)_{t \geq 0}$. In this sense, the process $(Y_t)_{t \geq 0}$ can be interpreted as an *information process*, reflecting the average dynamics of the information about the process $(X_t)_{t \geq 0}$.

It remains to define a new cost function such that for each policy the average costs of the process $(Y_t)_{t \geq 0}$ coincide with those of the process $(X_t)_{t \geq 0}$. In fact, an adequate choice is just given by the cost function $\tilde{C}(x,a,\tau)$ defined in (2.27) denoting the average costs within the time interval $[0,\tau)$ of hidden progress under the condition that the process $(X_t)_{t\geq 0}$ starts in state $x \in \mathcal{S}$ and action $a \in \mathcal{A}$ is chosen. For infinite lag times we set

$$\tilde{C}(x,a,\infty) = C(x,a,\infty) := \lim_{T\to\infty} \mathbb{E}_x^a \left( \frac{1}{T} \int_0^T c(X_s,a)\,ds \right). \tag{2.35}$$

Now we can state the analogy of the processes $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ with respect to the average costs for a given policy.

**Lemma 2.24.** *For each policy $u \in \mathcal{U}$ the freely observable process $(Y_t)_{t\geq 0}$ together with the cost function $\tilde{C}$ has the same expected average costs as the process $(X_t)_{t\geq 0}$, i.e. it holds*

$$\lim_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T \tilde{C}(Y_s, u(Y_s))\,ds \right) = \bar{J}(x,u) \tag{2.36}$$

*with $\bar{J}(x,u)$ given in* (2.24).

*Proof.* For ergodic dynamics and finite lag times we only need to show that $\tilde{\mu}$ (defined in (2.28)) is the equilibrium distribution for the process $(Y_t)_{t\geq 0}$, which simply follows from

$$\tilde{\mu} G_u = \tilde{\mu} \begin{pmatrix} \ddots & & 0 \\ & \frac{1}{\tau(x)} & \\ 0 & & \ddots \end{pmatrix} (P_u - I) = \frac{1}{\sum_{y\in\mathcal{S}} \tau(y)\mu(y)} (\mu P_u - \mu I) = 0$$

because $\mu P_u = \mu$. Then it holds

$$\lim_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T \tilde{C}(Y_s, u(Y_s))\,ds \right) = \sum_{y\in\mathcal{S}} \tilde{\mu}(y)\tilde{C}(y,u(y)) = \bar{J}(x,u) \quad \forall x \in \mathcal{S},$$

compare (2.29).

In the case of non ergodic dynamics we can carry out the same analysis for each of the communication classes which coincide for the two processes $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$. For those states which do not belong to any communication class (because they are either absorbing states or cannot be reached by any other state) it suffices to note that this property is also preserved when switching from one process to the other.

For $\tau(x) = \infty$ the process $(Y_t)_{t\geq 0}$ almost surely stays in $x$ such that

$$\begin{aligned} \lim_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T \tilde{C}(Y_s, u(Y_s))\,ds \right) &= \lim_{T\to\infty} \mathbb{E}_x^u \left( \frac{1}{T} \int_0^T \tilde{C}(x, a(x), \infty)\,ds \right) \\ &= \tilde{C}(x, a(x), \infty) \\ &= \bar{J}(x,u), \end{aligned}$$

compare (2.35) and (2.24).

For $X_0 = Y_0 = y \neq x$ with $\tau(y) < \infty$ but $\tau(x) = \infty$, both processes $(X_{t_j})_{j\in\mathbb{N}_0}$

and $(Y_t)_{t\geq 0}$ will reach $x$ almost surely after a finite amount of time (assuming that $x$ and $y$ communicate). As such a single finite time interval has no impact on the long-term average costs, this means for both processes that the long-term average costs starting in $y$ are given by the long-term average costs starting in $x$. If – in this situation – there exist several states $x \in \mathcal{S}$ with $\tau(x) = \infty$, both processes will reach one of them almost surely after a finite amount of time. Which one will be reached is random, but the probabilities coincide for both processes. The long-term average costs starting in $y$ are given by the weighted mean of the long-term average costs of the states $x$. □

Note that in (2.36) the second argument $u(Y_s)$ of the cost function $\tilde{C}$ is running in time, whereas in the definition (2.5) resp. (2.24) of $\bar{J}(x, u)$ it was evaluated at the beginning $t_j$ of a time interval.

**Example 2.25** (Two states)**.** *In order to get an impression of how the processes $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ evolve over time, we consider again the 2-state-example 2.11 from Section 2.2 with $\mathcal{S} = \{x_1, x_2\}$ and $\mathcal{A} = \{a_1, a_2\}$ as well as*

$$L_1 = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix}, \quad L_2 = \begin{pmatrix} -0.1 & 0.1 \\ 0.1 & -0.1 \end{pmatrix}$$

*and $c(x, a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$, where $c_{\mathcal{S}}(x_1) = 0$, $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_1) = 0$, $c_{\mathcal{A}}(a_2) = 2$. For the policy $u(x) = (a(x), \tau(x))$ we choose $a(x_1) = a_1$, $\tau(x_1) = 5$, $a(x_2) = a_2$, $\tau(x_2) = 2$.*
*The corresponding matrix $P_u$ is*

$$P_u = \begin{pmatrix} 0.9524 & 0.0476 \\ 0.1648 & 0.8352 \end{pmatrix},$$

*with the stationary distribution*

$$\mu = \begin{pmatrix} 0.7760 & 0.2240 \end{pmatrix},$$

*which delivers*

$$\tilde{\mu} = \begin{pmatrix} 0.8965 & 0.1035 \end{pmatrix}.$$

*The relevant values of the cost function $\tilde{C}$ are given by*

$$\tilde{C}(x_1, a_1, 5) = 0.4419, \quad \tilde{C}(x_2, a_2, 2) = 11.6210,$$

*and we get*

$$\eta_u = \sum_{x \in \mathcal{S}} \tilde{\mu}(x) \tilde{C}(x, u(x)) = 1.5989.$$

*Finally, the resulting generator $G_u$ of the process $(Y_t)_{t\geq 0}$ is given by*

$$G_u = \begin{pmatrix} -0.0095 & 0.0095 \\ 0.0824 & -0.0824 \end{pmatrix}.$$

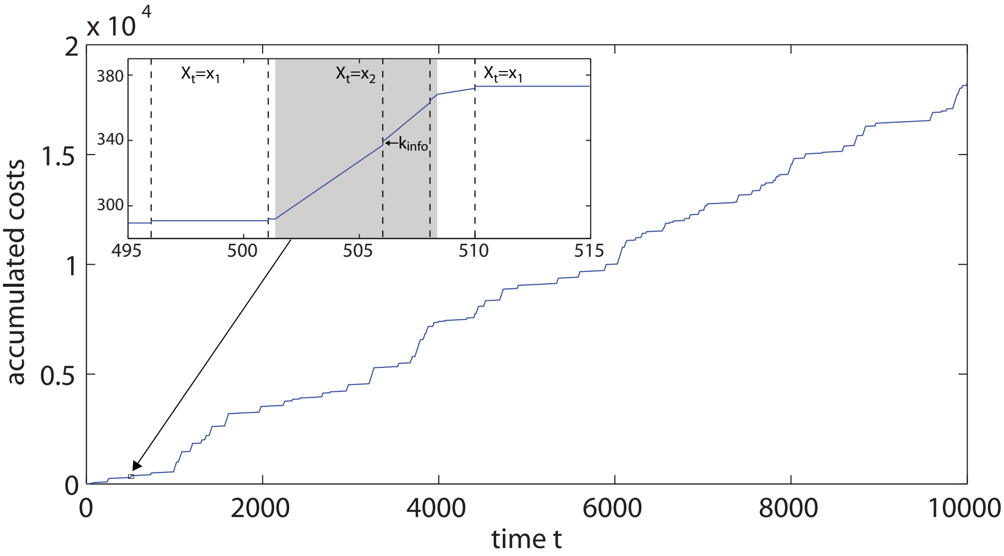

Figure 2.14: **Accumulated costs for $(X_t)_{t\geq 0}$.** Accumulated costs $J_t(x_1, u)$ defined in (2.37) up to time $t \geq 0$ for a trajectory of the Markov decision process $(X_t)_{t\geq 0}$ with information costs $k_{\text{info}}$ described in Example 2.25. The long-term asymptotics are given by $\eta_u \cdot t = 1.5989 \cdot t$. The detail shows the cost increase for the time period $t \in [495, 515]$. The dashed lines are located at the observation times $t_j$; at each of these observation times the costs increase instantaneously by $k_{\text{info}}$. The gray area indicates the period of time where the process $(X_t)_{t\geq 0}$ is in state $x_2$, while it is in state $x_1$ at all other times.

*Figure 2.14 shows the accumulated costs*

$$J_t(x_1, u) := \sum_{j=0}^{n(t)-1} \left( \int_{t_j}^{t_{j+1}} c(X_s, a(X_{t_j}))\, ds + k_{\text{info}} \right) + \int_{t_{n(t)}}^{t} c(X_s, a(X_{t_{n(t)}}))\, ds, \tag{2.37}$$

$$n(t) := \max\{j \in \mathbb{N} : t_j \leq t\},$$

*up to time $t > 0$ for a realization of the process $(X_t)_{t\geq 0}$ starting in $X_0 = x_1$ and given control $u$. It contains a detailed view for the time interval $t \in [495, 515]$ which illustrates the structure of cost increase for the situation of information costs.*

*We can see that after the observation at time $t = 501$ the process switches to state $x_2$ which leads to a higher increase in the costs. At time $t = 506$ this switch is observed and the action is adapted. Now the more expensive action $a_2$ leads again to a higher increase in the costs, but at the same time accelerates the return to state $x_1$ which happens at $t \approx 508.4$. At time $t = 510$ this is realized and action $a_2$ is replaced by action $a_1$. Every observation leads to a jump in the accumulated costs of size $k_{\text{info}} = 1$.*

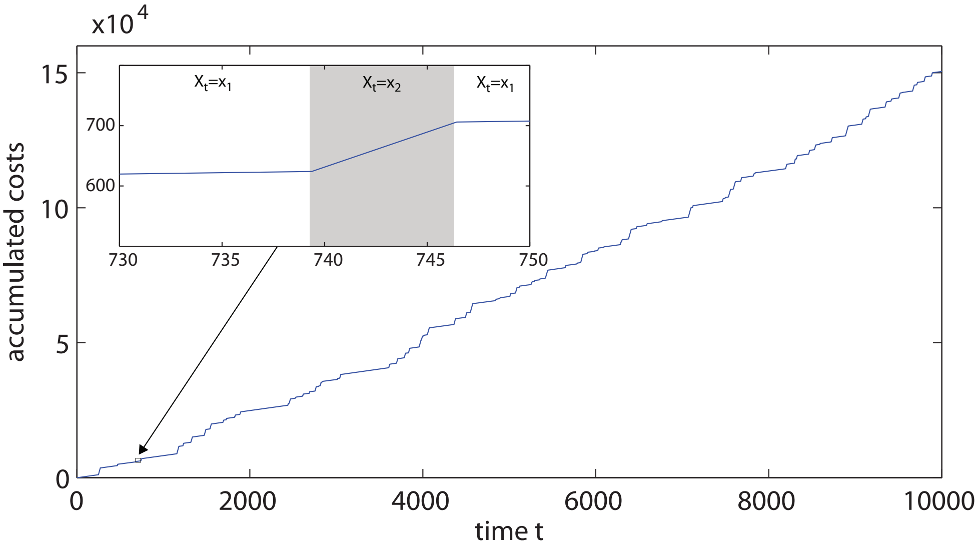

Figure 2.15: **Accumulated costs for $(\mathbf{Y}_t)_{t\geq 0}$.** Accumulated costs $\tilde{J}_t(x_1, u)$ defined in (2.38) up to time $t \geq 0$ for a trajectory of the equivalent freely observable Markov decision process $(Y_t)_{t\geq 0}$ given Example 2.25. Like for the process $(X_t)_{t\geq 0}$ the long-term asymptotics are given by $\eta_u \cdot t = 1.5989 \cdot t$. The detail shows the cost increase for the time period $t \in [730, 750]$. Again, the gray area indicates the period of time where the process $(X_t)_{t\geq 0}$ is in state $x_2$, while it is in state $x_1$ at all other times. Observations are continuous (and cost-free) over time and the action is immediately adapted after a switch of the state occurs.

*Equivalent graphs are given for the freely observable process $(Y_t)_{t\geq 0}$: Figure 2.15 shows the accumulated costs*

$$\tilde{J}_t(x_1, u) := \int_0^t \tilde{C}(Y_s, u(Y_s))\, ds \tag{2.38}$$

*up to time $t > 0$ for a realization of the process $(Y_t)_{t\geq 0}$ starting in $Y_0 = x_1$ and given control $u$. It contains the details for the time interval $t \in [730, 750]$.*

*For the process $(Y_t)_{t\geq 0}$ a change in the state is followed by an instantaneous change in the action, and so there are only two increase rates: A low increase $\tilde{C}(x_1, a_1, 5)$ (induced by the information costs $k_{\text{info}}$ that are included in $\tilde{C}$) when the process is in state $x_1$, and a high increase $\tilde{C}(x_2, a_2, 2)$ (induced by positive action- and state costs and $k_{\text{info}}$) when the process is in state $x_2$.*

The preceding analysis permits a straightforward transfer of the results presented in Section 1.1.2 to the new setting of Markov control with information costs and policies with finite lag times. The case of infinite lag times will require a separate investigation as it breaches the ergodicity condition. Before turning to optimal policies, we will now analyze the average costs for an arbitrary policy.

**Average costs for a given policy $u \in \mathcal{U}$**

For a policy $u \in \mathcal{U}$ with **finite lag times**, i.e. $\tau(x) < \infty \ \forall x \in \mathcal{S}$, the process $(Y_t)_{t\geq 0}$ (driven by the generator $G_u$, see (2.33)) is ergodic such that the average costs are given by a constant $\eta_u$. Remember that we assumed the state space to be finite, such that Lemma 1.13 can directly be applied replacing the cost function $c$ by $\tilde{C}$ and the generator $L_u$ by $G_u$.

**Lemma 2.26.** *Consider the Markov control model with information costs* (2.1)*.*

*a) Given a policy $u(x) = (a(x), \tau(x))$ with $\tau(x) < \infty \ \forall x \in \mathcal{S}$, there exists a function $v : \mathcal{S} \to \mathbb{R}$ such that the corresponding constant $\eta_u$ of long-run expected average costs (compare* (2.29)*) fulfills the equation*

$$\eta_u \quad = \quad \tilde{C}(x, a(x), \tau(x)) + \sum_{y \in \mathcal{S}} G_{a(x),\tau(x)}(x, y) v(y) \quad \forall x \in \mathcal{S} \tag{2.39}$$

*with $\tilde{C}$ resp. $G_{a,\tau}$ defined in* (2.27) *resp.* (2.30)*.*

*b) The constant $\eta_u$ is uniquely determined by* (2.39) *and coincides with the first component of the vector*

$$(E - G_u)^{-1}\tilde{C}_u,$$

*where*

$$E := \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|},$$

*and $\tilde{C}_u(x) = \tilde{C}(x, a(x), \tau(x))$.*

*Proof.* This directly follows from Lemma 1.13 and Lemma 2.24. □

**Lemma 2.27.** *Let $u \in \mathcal{U}$ be a given policy with finite lag times. It holds:*

*a) If there exists a constant $g \geq 0$ and a function $v : \mathcal{S} \to \mathbb{R}$ such that*

$$g \geq \tilde{C}(x, a(x), \tau(x)) + \sum_{y \in \mathcal{S}} G_{a(x),\tau(x)}(x, y) v(y) \quad \forall x \in \mathcal{S}, \tag{2.40}$$

*then $g \geq \eta_u$.*

*b) If there exists a constant $g \geq 0$ and a function $v : \mathcal{S} \to \mathbb{R}$ such that*

$$g \leq \tilde{C}(x, a(x), \tau(x)) + \sum_{y \in \mathcal{S}} G_{a(x),\tau(x)}(x, y) v(y) \quad \forall x \in \mathcal{S}, \tag{2.41}$$

*then $g \leq \eta_u$.*

*Proof.* This is just an analogue of Lemma 1.14. □

In the case of **infinite lag times** the long-term average costs might depend on the initial state: Imagine a policy $u \in \mathcal{U}$ fulfilling $\tau(x) = \tau(y) = \infty$ but $a(x) \neq a(y)$ for two states $x \neq y$. That is, starting in $X_0 = x \in \mathcal{S}$ the process $(X_t)_{t \geq 0}$ will forever be steered by $L_{a(x)}$ and produce costs according to $c(\cdot, a(x))$, while for $X_0 = y \in \mathcal{S}$ action $a(y)$ determines the remaining dynamics. Generally, it will hold $C(x, a(x), \infty) \neq C(y, a(y), \infty)$ for the corresponding long-term average costs, and with it

$$\bar{J}(x, u) \neq \bar{J}(y, u).$$

This is consistent with the choice of the generator $G_{a,\infty}$, compare (2.31): Its zero entries refer to the fact that for the equivalent control process $(Y_t)_{t \geq 0}$ the considered states $x, y$ are absorbing. In other words, given infinite lag times, the dynamics of $(Y_t)_{t \geq 0}$ are not ergodic and so the long-term average costs are in general not given by a constant. If in this situation there exists another state $z \in \mathcal{S}$ with finite lag time $\tau(z) < \infty$, the process will (after starting in $z$) reach one of the two states $x, y$ after some random period of time, and thus the expected long-term average costs will be a weighed average of $\bar{J}(x, u)$ and $\bar{J}(y, u)$.
We can conclude that the statements of Lemma 2.26 and Lemma 2.27 have no direct analogue for infinite lag times. Instead, in this case the calculation of the long-term average costs requires a separate analysis for each of the given states.

Fortunately, such a separate analysis will be redundant in the case of optimal policies. The described situation with $\bar{J}(x, u) \neq \bar{J}(y, u)$ naturally excludes the referring policy $u$ from being optimal as it either holds $\bar{J}(x, u) > \bar{J}(y, u)$ or $\bar{J}(x, u) < \bar{J}(y, u)$. In the first case the long-term average costs when starting in $x$ could be decreased by choosing action $a(y)$ instead of the given $a(x)$ which would lead to a policy $\tilde{u}$ with

$$\bar{J}(x, \tilde{u}) = C(x, a(y), \infty) \overset{(*)}{=} C(y, a(y), \infty) = \bar{J}(y, u) < \bar{J}(x, u).$$

The second equality $(*)$ is due to the fact that the underling dynamics are assumed to be ergodic such that the long-term average costs in the case of infinite lag times only depend on the action but not on the initial state. In the case of $\bar{J}(x, u) < \bar{J}(y, u)$ we simply interchange the roles of $x$ and $y$ in order to show that the given policy $u$ cannot be optimal. By this argumentation we can see that the long-term average costs of an *optimal* policy actually will be given by a constant.

In the following, we will use this insight for the analysis of optimal policies. Again, we can directly transfer the results from the original setting to the new model with information costs.

### Optimal policy and value function

Our aim is now to characterize the constant $\eta^*$ of optimal average costs,

$$\eta^* := \inf_{u \in \mathcal{U}} \eta_u, \tag{2.42}$$

as well as the optimal policy (if existent) for the Markov decision process with information costs. In fact, we simply tend to reformulate Theorem 1.15 (the Bellman equation) and Theorem 1.16 (characterization of the optimal policy) of Section 1.1.2 in terms of the new setting. The only crucial step is to check Assumption 1.11 which was used in Section 1.1.2 to prove both theorems. It suggests that the set of available actions is compact for each state $x \in \mathcal{S}$ and that both the cost function and the generators are continuous in the action parameter. (As we consider a finite state space we can ignore the third part of Assumption 1.11.) In the new setting, the action parameter is of the form $(a, \tau)$ and the set of available actions for any state $x \in \mathcal{S}$ is given by $\mathcal{A} \times (0, \infty]$. As in the case of discounted costs (compare page 54) we can state that, by $k_{\text{info}} > 0$, it holds

$$\lim_{\tau(x) \to 0} \bar{J}(x, u) = \infty \quad \forall x \in \mathcal{S}.$$

Again this means that we can find a lower bound $\varepsilon > 0$ for the optimal lag times such that the relevant set of actions can be restricted to $\mathcal{A} \times [\varepsilon, \infty]$.
Therefore, the compactness condition is naturally fulfilled as long as $\mathcal{A}$ is compact. It remains to note that the new cost function $\tilde{C}$ defined in (2.27) and the generators $G_{a,\tau}$ defined in (2.30) are all continuous in $\tau$ (compare the statement in (2.32) and the definition in (2.35) for $\tau = \infty$), such that the continuity condition only needs to be checked for $\mathcal{A}$. We will write the function $\tilde{C}$ as a composition of $C$ and $\frac{k_{\text{info}}}{\tau}$ in order to directly note the parameter $k_{\text{info}}$, setting $\frac{k_{\text{info}}}{\infty} := 0$ for infinite lag times. As the second component $\frac{k_{\text{info}}}{\tau}$ does not depend on $a$, the continuity condition with respect to $a \in \mathcal{A}$ only concerns the function $C$.

Hence, the formulation of the average-cost optimality equation for Markov decision processes with information costs is the following.

**Theorem 2.28** (Average cost optimality equation/Bellman equation)**.**
*Suppose that $\mathcal{A}$ is a compact set and that for all $x, y \in \mathcal{S}$ and $\tau \in (0, \infty]$ the functions $C(x, a, \tau)$ and $G_{a,\tau}(x, y)$ are continuous in $a \in \mathcal{A}$. Then there exists a function $v^* : \mathcal{S} \to \mathbb{R}$ and a constant $g \geq 0$ satisfying*

$$g = \inf_{a \in \mathcal{A}, \tau \in (0,\infty]} \left\{ C(x, a, \tau) + \frac{k_{\text{info}}}{\tau} + \sum_{y \in \mathcal{S}} G_{a,\tau}(x, y) v^*(y) \right\} \quad \forall x \in \mathcal{S}. \tag{2.43}$$

*The constant $g$ coincides with the optimal average costs $\eta^*$.*

*Proof.* By the proceeding argumentation this is the adequate analogue to Theorem 1.15. □

**Theorem 2.29.** *Suppose that there exists a policy $u^*(x) = (a^*(x), \tau^*(x)) \in \mathcal{U}$ which attains the minimum in the Bellman equation* (2.43)*, i.e.*

$$\begin{aligned} g &= C\left(x, a^*(x), \tau^*(x)\right) + \frac{k_{\text{info}}}{\tau^*(x)} + \sum_{y\in\mathcal{S}} G_{a^*(x),\tau^*(x)}(x,y)v^*(y) \\ &= \min_{a\in\mathcal{A},\tau\in(0,\infty]} \left\{ C(x,a,\tau) + \frac{k_{\text{info}}}{\tau} + \sum_{y\in\mathcal{S}} G_{a,\tau}(x,y)v^*(y) \right\} \quad \forall x \in \mathcal{S}. \end{aligned} \tag{2.44}$$

*Then $u^*$ is average-cost optimal and it holds $\eta^* = \bar{J}(x, u^*) = g$ for all $x \in \mathcal{S}$.*

*Proof.* This is the analogue to Theorem 1.16. □

In Section 1.1.2 we stated that Assumption 1.11 guarantees the existence of an optimal policy for the average-cost criterion. We just noted that – within the setting of Markov control with information costs – the compactness and continuity conditions of Assumption 1.11 only have to be checked for $\mathcal{A}$, whereas they are naturally fulfilled for the lag time parameter $\tau$. That is, as long as $\mathcal{A}$ is compact and $\tilde{C}$ and $G_{a,\tau}$ are continuous with respect to $a \in \mathcal{A}$, the existence of an optimal policy can be taken for granted. This is especially fulfilled for finite $\mathcal{A}$.

For a numerical calculation of the optimal policy for the process $(X_t)_{t\geq 0}$ we can apply the average costs value iteration and policy iteration algorithms of Section 1.2 to the equivalent completely observable process $(Y_t)_{t\geq 0}$ deduced on page 81. As in the case of discounted costs, the complexity of the problem dramatically increases due to the lag time parameter $\tau$. The idea how to deal with this complexity in the case of average costs agrees with the one given for the discounted-cost criterion on 55. Also the analysis of numerical alternatives is the same. Again we can state that a policy iteration combined with a discretization of the lag time parameter $\tau$ delivers a satisfying optimization method.

**Example 2.25 (cont.)** *The optimal policy for the 2-state-example 2.25 with different cost parameters is given in Table 2.2. As in the case of discounted costs, the "good" state $x_1$ causes the application of $a_1$, while the "bad" state $x_2$ requires the application of the more expensive action $a_2$ for a short time. An increase of the information costs induces an increase in the lag times (compare the first two rows). Higher action costs $c_{\mathcal{A}}(a_2)$, however, reduce the time for its application (compare row 1 and 3).*

*Comparing the results to the case of discounted costs, see Table 2.1, we can see that the optimal time intervals are much shorter for the case of average costs. An explanation for this relation will be given in Section 2.4.*

*The value of optimal costs $\eta$ can not be compared to the value function of the discounted-cost problem as one would compare costs per unit of time with total discounted costs.*

**Example 2.30** (Three states)**.** *For the discounted-cost criterion we analyzed, given the 2-state-model, how an additional intermediate state influences the optimal policy,*

| $c_{\mathcal{S}}(x_2)$ | $c_{\mathcal{A}}(a_2)$ | $k_{\text{info}}$ | $a^*(x_1)$ | $a^*(x_2)$ | $\tau^*(x_1)$ | $\tau^*(x_2)$ | $\eta^*$ |
|---|---|---|---|---|---|---|---|
| 10 | 2 | 1 | $a_1$ | $a_2$ | 5.3 | 1.3 | 1.59 |
| 10 | 2 | 2 | $a_1$ | $a_2$ | 7.7 | 1.8 | 1.79 |
| 10 | 3 | 1 | $a_1$ | $a_2$ | 5.4 | 1.2 | 1.68 |

Table 2.2: **Parameter dependent optimal policy and optimal average costs for the 2-state-example.** This table shows the optimal policy for Example 2.25 and different values of $c_{\mathcal{S}}(x_2)$, $c_{\mathcal{A}}(a_2)$ and $k_{\text{info}}$, given the average-cost criterion. In state $x_1$ (resp. $x_2$) one has to choose action $a^*(x_1)$ (resp. $a^*(x_2)$) for a time period $\tau^*(x_1)$ (resp. $\tau^*(x_2)$) which results in optimal costs $\eta^*$ (independent of the state).

*see Example 2.12. Now we do the same for the average-cost criterion, see Table 2.3. Comparing these results to the ones of the 2-state-example given in Table 2.2 we can observe slightly better optimal costs. The lag times in state $x_1$ and $x_2$ increase, and the intermediate state $x_I$ has a relatively short waiting time.*

| $c_{\mathcal{S}}(x_2)$ | $c_{\mathcal{A}}(a_2)$ | $k_{\text{info}}$ | $a^*(x_1)$ | $a^*(x_I)$ | $a^*(x_2)$ | $\tau^*(x_1)$ | $\tau^*(x_I)$ | $\tau^*(x_2)$ | $\eta^*$ |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 2 | 1 | $a_1$ | $a_1$ | $a_2$ | 8.8 | 3.0 | 1.4 | 1.52 |
| 10 | 2 | 2 | $a_1$ | $a_1$ | $a_2$ | 11.7 | 4.3 | 2.2 | 1.70 |
| 10 | 3 | 1 | $a_1$ | $a_1$ | $a_2$ | 8.9 | 3.0 | 1.3 | 1.62 |

Table 2.3: **Parameter dependent optimal policy and optimal average costs for the 3-state-example.** This table shows the optimal policy for Example 2.30 and different values of $c_{\mathcal{S}}(x_2)$, $c_{\mathcal{A}}(a_2)$ and $k_{\text{info}}$, given the average-cost criterion. In state $x_1$ (resp. $I$ resp. $x_2$) one has to choose action $a^*(x_1)$ (resp. $a^*(I)$ resp. $a^*(x_2)$) for a time period $\tau^*(x_1)$ (resp. $\tau^*(I)$ resp. $\tau^*(x_2)$) which results in optimal costs $\eta^*$ (independent of the state).

*This is exactly what one would expect: As direct transitions from the "good" state $x_1$ to the costly state $x_2$ are not possible, state $x_1$ can be interpreted as a "safe" area which does not require frequent testing. However, finding the process in the intermediate state $x_I$ makes a soon following transition to $x_2$ more likely, such that the optimal control policy proposes to quickly make a test again. The additional information resulting from the splitting of the original state $x_1$ into a safe area and a transition region leads to better optimal costs.*

The example of a controlled population (compare Example 2.13) will not be considered here: As the model contains an absorbing state, the average-cost criterion is not suited to evaluate the dynamics.

### 2.3.2 Cost Splitting for Average Costs

As in the case of discounted costs, we can ask ourselves how the optimal costs $\eta^*$ split up into components of information costs, action costs and state costs. Again, we will profit from the formulation of the equivalent freely observable process $(Y_t)_{t\geq 0}$, see page 81, and express these components – as for finite lag times – in terms of the equilibrium distribution $\tilde{\mu}$ of the generator $G_{u^*}$ defined in (2.33), where $u^*$ is the optimal policy. The case of infinite lag times will be analyzed separately.

Assuming the process to be ergodic and the optimal lag times $\tau^*(x)$ to be finite, we can write

$$\eta^* = \lim_{T\to\infty} \mathbb{E}_x \left( \frac{1}{T} \int_0^T \tilde{C}(Y_s, u^*(Y_s))\, ds \right) = \sum_{y\in\mathcal{S}} \tilde{\mu}(y) \cdot \tilde{C}(y, u^*(y)) \tag{2.45}$$

for the average costs under optimal control. Let the cost function $c$ be of the form $c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$ and write

$$\begin{aligned} \tilde{C}(x,a,\tau) &= \mathbb{E}_x^a \left( \frac{1}{\tau} \int_0^\tau c_{\mathcal{S}}(X_s) + c_{\mathcal{A}}(a)\, ds \right) + \frac{k_{\text{info}}}{\tau} \\ &= \mathbb{E}_x^a \left( \frac{1}{\tau} \int_0^\tau c_{\mathcal{S}}(X_s)\, ds \right) + c_{\mathcal{A}}(a) + \frac{k_{\text{info}}}{\tau}. \end{aligned}$$

Now it is easy to split the total average costs $\eta$ of an arbitrary policy with finite lag times into its components: Set

$$\eta = \eta_{\text{info}} + \eta_{\mathcal{A}} + \eta_{\mathcal{S}}$$

with

$$\eta_{\text{info}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot \frac{k_{\text{info}}}{\tau(x)},$$

$$\eta_{\mathcal{A}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot c_{\mathcal{A}}(a(x)),$$

$$\eta_{\mathcal{S}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot C_{\mathcal{S}}(x, a(x), \tau(x)),$$

and

$$C_{\mathcal{S}}(x,a,\tau) := \mathbb{E}_x^a \left( \frac{1}{\tau} \int_0^\tau c_{\mathcal{S}}\left(X_s\right)\, ds \right). \tag{2.46}$$

In Lemma 2.26 we stated that the total average costs $\eta$ are given by the first entry of the vector

$$(E - G_u)^{-1} \tilde{C}_u$$

where

$$E := \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix}.$$

Equivalent formulas hold for the components $\eta_{\text{info}}$, $\eta_{\mathcal{A}}$ and $\eta_{\mathcal{S}}$:

**Lemma 2.31** (Information costs). *The average information costs $\eta_{\text{info}}$ for the optimal policy $u^*$ are given by the first entry of the vector*

$$k_{\text{info}} \cdot (E - G_{u^*})^{-1} \begin{pmatrix} \vdots \\ \frac{1}{\tau^*(x)} \\ \vdots \end{pmatrix}.$$

**Lemma 2.32** (Action costs). *The average action costs $\eta_{\mathcal{A}}$ for the optimal policy $u^*$ are given by the first entry of the vector*

$$(E - G_{u^*})^{-1} c^*_{\mathcal{A}}$$

*with $c^*_{\mathcal{A}}(x) := c_{\mathcal{A}}(a^*(x))$ for all $x \in \mathcal{S}$.*

**Lemma 2.33** (State costs). *The average state costs $\eta_{\mathcal{S}}$ for the optimal policy $u^*$ are given by the first entry of the vector*

$$(E - G_{u^*})^{-1} C^*_{\mathcal{S}}$$

*with $C^*_{\mathcal{S}}(x) := C_{\mathcal{S}}(x, a^*(x), \tau^*(x))$ for all $x \in \mathcal{S}$, compare* (2.46).

All three statements follow directly from Lemma 2.26 when setting for each case two of the three components $c_{\mathcal{S}}$, $c_{\mathcal{A}}$ and $k_{\text{info}}$ in $\tilde{C}$ to zero, which results in a cost functional containing only information-, action- or state costs, respectively.
As in the case of discounted costs, all the statements can be formulated in an equivalent way for any non optimal policy $u \neq u^*$ with finite lag times.

In the case of infinite lag times the information costs vanish since no state tests are made. (This even holds for a state with finite lag time as long as it communicates with another state with infinite lag time because – in the long run – this state will almost surely be reached, and so the number of tests will stay finite.)
The action costs and the state costs possibly depend on the state: The action costs of a state $x \in \mathcal{S}$ with infinite lag time $\tau(x) = \infty$ are given by

$$\bar{J}_{\mathcal{A}}(x, u) = c_{\mathcal{A}}(a(x));$$

and the corresponding state costs are given by

$$\bar{J}_{\mathcal{S}}(x, u) = \lim_{T \to \infty} \mathbb{E}_x^{a(x)} \left( \frac{1}{T} \int_0^T c_{\mathcal{S}}(X_s)\, ds \right).$$

### Comparison to the original Markov control problem

Let us use the presented cost splitting of the optimal average costs in order to compare the given Markov control problem with information costs to the original problem. Figure 2.16 shows the optimal average costs for both settings applied to the 2-state-example 2.25. Naturally, the optimal value of the Markov control problem with information costs exceeds the one of the original control model. The same is true when comparing only the net costs $\eta_{\text{net}} = \eta_{\mathcal{S}} + \eta_{\mathcal{A}}$ (blue and green area) to the optimal costs of the original model. This relation holds in general and not only for the considered example. The justifying argumentation is the same as for the discounted-cost criterion, compare Section 2.2.2: In order to guarantee an (overall) optimal control, the points in time where the action is adapted have to coincide with the jumping times of the process. Otherwise, there will be periods in time where the influence on the process is not optimal.

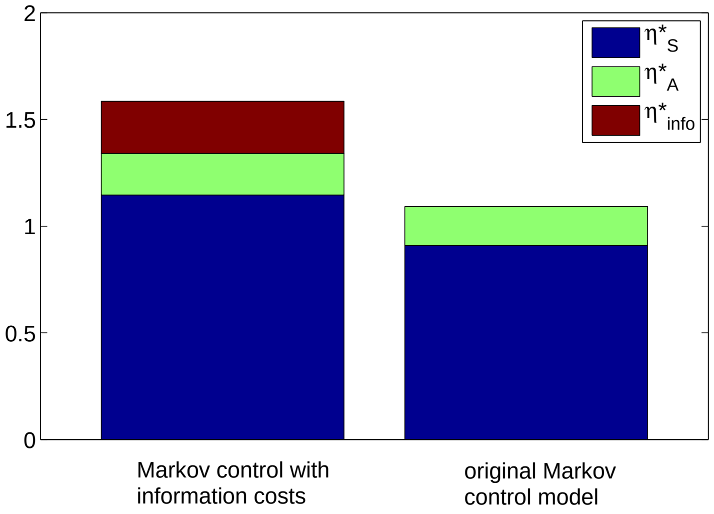

Figure 2.16: **Cost splitting for the 2-state-example.** The constant of optimal average costs $\eta^*$ defined in (2.42) for Example 2.25 is divided into its components $\eta^*_{\mathcal{S}}$, $\eta^*_{\mathcal{A}}$ and $\eta^*_{\text{info}}$ and compared to the free information case where the splitting is given by $\eta^*_{\mathcal{S}} + \eta^*_{\mathcal{A}}$. The respective cost parameters are given by $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_2) = 2$ and $k_{\text{info}} = 1$, compare first row of Table 2.2.

### 2.3.3 Monotonicity and Continuity with respect to $k_{\text{info}}$

We continue to consider ergodic dynamics such that the value function $\bar{V}$ of optimal average costs is given by a constant $\eta^*$. (Remember that by the consideration from page 88 the optimal long-term average costs have to be constant even if the optimal lag times are infinite.) Given the insight into the characteristics of this constant $\eta^*$ and its different components, we now turn to the analysis of the central parameter $k_{\text{info}}$ and its influence on the optimal costs resp. the optimal lag times. Again we ask the questions given in Section 2.2.3: Is the constant of optimal average costs $\eta^*$ of optimal costs monotone and continuous with respect to $k_{\text{info}}$? And how do the optimal lag times depend on $k_{\text{info}}$?

### Monotonicity and continuity of $\boldsymbol{\eta^*}$ with respect to $\boldsymbol{k_{\text{info}}}$

**Lemma 2.34** (Monotonicity of $\eta^*$ with respect to $k_{\text{info}}$)**.** *Let $\eta^* = \inf_{u \in \mathcal{U}} \bar{J}(x,u)$ be the optimal average costs of a given control problem with information cost parameter $k_{\text{info}}$. Changing the information cost parameter to $\tilde{k}_{\text{info}}$ with $\tilde{k}_{\text{info}} < k_{\text{info}}$ results in optimal average costs $\tilde{\eta}^*$ with*

$$\tilde{\eta}^* \leq \eta^*.$$

*Proof.* The argumentation is analogous to the one in Lemma 2.18: Let $u^*$ be the optimal policy with respect to the parameter $k_{\text{info}}$, i.e. it holds $\eta^* = \bar{J}(x,u^*)$ for all $x$. For a fixed policy, the cost functional $\bar{J}$, defined in (2.24), is obviously monotone in $k_{\text{info}}$. This yields

$$\tilde{J}(x,u^*) \leq \bar{J}(x,u^*),$$

where $\tilde{J}$ is the cost functional for the parameter $\tilde{k}_{\text{info}}$. By $\tilde{\eta}^* \leq \tilde{J}(x,u^*)$ $\forall x \in \mathcal{S}$ the statement follows. □

**Theorem 2.35** (Continuity of $\eta^*$ with respect to $k_{\text{info}}$ at $k_{\text{info}} > 0$)**.** *The quantity $\eta^*$ of optimal average costs is continuous in $k_{\text{info}} > 0$.*

*Proof.* Analogous to the proof of Theorem 2.19. □

**Theorem 2.36** (Continuity of $\eta^*$ with respect to $k_{\text{info}}$ at $k_{\text{info}} = 0$)**.** *Given a Markov control model with information costs, let $\eta^*_{k_{\text{info}}}$ denote the optimal average costs depending on the parameter $k_{\text{info}}$. Let $\eta_0$ denote the optimal average costs of the corresponding original Markov Control model (without information costs). It holds*

$$\eta^*_{k_{\text{info}}} \overset{k_{\text{info}} \to 0}{\longrightarrow} \eta_0.$$

Figure 2.17 shows the optimal average costs $\eta^*$ as a function of $k_{\text{info}}$ for the 2-state-example 2.25.

*Proof of Theorem 2.36.* Again, the argumentation is analogous to the case of discounted costs, compare Theorem 2.20. We consider the optimal policy $a_0$ of the original Markov control problem. Defining $c_{a_0}(x) := c(x, a_0(x))$ and $L_{a_0}(x,y) := L_{a_0(x)}(x,y)$, $x, y \in \mathcal{S}$, we get

$$\eta_0 = c_{a_0}(x) + (L_{a_0}v)(x) \quad \forall x \in \mathcal{S},$$

where $v$ is a suitable function on $\mathcal{S}$. Given $k_{\text{info}} > 0$, set $\tau(x) = \tau^* = \sqrt{k_{\text{info}}}$ for all $x \in \mathcal{S}$ and consider the policy $u(x) = (a_0(x), \tau^*)$ which is in general not optimal. According to Lemma 2.26 the average costs $\eta_{k_{\text{info}}}$ induced by this policy $u$ are given by the first component of

$$v_{k_{\text{info}}} := \left[ \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} - G_u \right]^{-1} \tilde{C}_u,$$

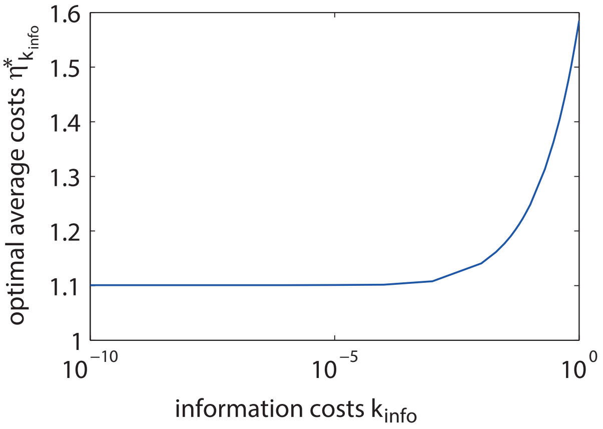

Figure 2.17: **$k_{\text{info}}$ vs. optimal average costs for the 2-state-example.** The constant of optimal average costs $\eta^*_{k_{\text{info}}}$ for Example 2.25 depending on the information costs $k_{\text{info}}$ in a logarithmic scale. For $k_{\text{info}} = 1$ the values agree with the one given in Table 2.2 ($\eta^* = 1.59$). For $k_{\text{info}} = 10^{-10}$ the value is close to the optimal costs of the original model ($\eta^* \approx 1.09$).

where, for the given policy $u$,

$$\tilde{C}_u(x) = \tilde{C}(x, a_0(x), \tau^*) = \mathbb{E}_x^{a_0(x)}\left(\frac{1}{\tau^*}\int_0^{\tau^*} c(X_s, a_0(x))\,ds\right) + \frac{k_{\text{info}}}{\tau^*}.$$

We analyze the vector $v_{k_{\text{info}}}$ for vanishing information costs: From $k_{\text{info}} \to 0$ it follows $\tau^* \to 0$ and

$$\frac{k_{\text{info}}}{\tau^*} = \frac{k_{\text{info}}}{\sqrt{k_{\text{info}}}} \to 0.$$

For $x \neq y$ it holds

$$\begin{aligned}
G_u(x,y) &= \frac{1}{\tau^*}e^{L_{a_0(x)}\tau^*}(x,y)\\
&= \frac{1}{\tau^*}\left[I + L_{a_0(x)}\tau^* + \frac{1}{2}L^2_{a_0(x)}(\tau^*)^2 + ...\right](x,y)\\
&\overset{I(x,y)=0}{=} \frac{1}{\tau^*}\left[L_{a_0(x)}\tau^* + \frac{1}{2}L^2_{a_0(x)}(\tau^*)^2 + ...\right](x,y)\\
&\xrightarrow{\tau^*\to 0} L_{a_0(x)}(x,y),
\end{aligned}$$

while

$$\begin{aligned}
G_u(x,x) &= -\frac{1}{\tau^*}\left(1 - e^{L_{a_0(x)}\tau^*}(x,x)\right)\\
&= -\frac{1}{\tau^*}\left(1 - \left[I + L_{a_0(x)}\tau^* + \frac{1}{2}L^2_{a_0(x)}(\tau^*)^2 + ...\right](x,x)\right)\\
&\overset{I(x,x)=1}{=} \frac{1}{\tau^*}\left[L_{a_0(x)}\tau^* + \frac{1}{2}L^2_{a_0(x)}(\tau^*)^2 + ...\right](x,x)\\
&\xrightarrow{\tau^*\to 0} L_{a_0(x)}(x,x),
\end{aligned}$$

and so we get $G_u \overset{\tau^* \to 0}{\longrightarrow} L_{a_0}$. Next, using $c_a(x) := c(x,a)$, we have

$$\begin{aligned}
&\mathbb{E}_x^u\left(\frac{1}{\tau^*}\int_0^{\tau^*} c(X_s, a_0(x))\,ds\right) \\
= \quad & \frac{1}{\tau^*}\int_0^{\tau^*}\left(e^{L_{a_0(x)}s}c_{a_0(x)}\right)(x)\,ds \\
= \quad & \frac{1}{\tau^*}\int_0^{\tau^*}\left(\left[I + L_{a_0(x)}s + \frac{1}{2}(L_{a_0(x)})^2 s^2 + ...\right]c_{a_0(x)}\right)(x)\,ds \\
= \quad & \frac{1}{\tau^*}\left[c_{a_0(x)}s + \frac{1}{2}(L_{a_0(x)}c_{a_0(x)})(x)s^2 + ...\right]_0^{\tau^*} \\
= \quad & c_{a_0(x)} + \frac{1}{2}(L_{a_0(x)}c_{a_0(x)})(x)\tau^* + ... \\
\overset{\tau^* \to 0}{\longrightarrow} \quad & c_{a_0(x)}(x).
\end{aligned}$$

Putting everything together we get

$$v_{k_{\text{info}}} \overset{k_{\text{info}} \to 0}{\longrightarrow} \left[\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} - L_{a_0}\right]^{-1} c_{a_0},$$

which, by Lemma 1.13, delivers $\eta_0$ in its first component, i.e. it holds $v_{k_{\text{info}}}(1) \overset{k_{\text{info}} \to 0}{\longrightarrow} \eta_0$. Noting that $\eta_0 \leq \eta^*_{k_{\text{info}}} \leq v_{k_{\text{info}}}(1) = \eta_{k_{\text{info}}}$ completes the proof. □

### Connection between $k_{\text{info}}$ and $\tau^*$

The analogy of the results for discounted and average costs persists when analyzing the connection between the information costs $k_{\text{info}}$ and the optimal lag times $\tau^*(x)$. We can even use the same 3-state-example (see Example 2.21) to show that $\tau^*(x)$ is in general not monotone and/or continuous in $k_{\text{info}}$. Figure 2.18 shows the optimal lag times $\tau^*(x)$ as functions of $k_{\text{info}}$ for the average-cost criterion. The corresponding function of optimal average costs is given in Figure 2.19.

The graph in Figure 2.18 looks similar to the one for discounted costs, compare Figure 2.9: For the states $x_1$ and $x_2$ the optimal lag time is monotone and continuous in $k_{\text{info}}$, while for the intermediate state $x_I$ there is a breaking point which is situated at $k_{\text{info}} \approx 0.12$. Again, this critical value of $k_{\text{info}}$ is linked to a switch in the optimal choice of action. For $k_{\text{info}} \leq 0.12$ the optimal action for state $x_I$ is $a^*(x_I) = a_1$, while for $k_{\text{info}} > 0.12$ it holds $a^*(x_I) = a_2$.

As before, this structure can be explained by the effect that both actions have on the process in state $x_I$: In the short run, action $a_1$ is preferred because it is free of charge ($c(x_I, a_1) = 0$). However, given $a_1$, the process is more likely to switch next to the "bad" state $x_2$, such that a soon following test and control adaption is required. If the test is too expensive ($k_{\text{info}} > 0.12$) it is favorable to choose the "safe" (but more expensive) action $a_2$ in order to push the process towards the "good" state $x_1$.

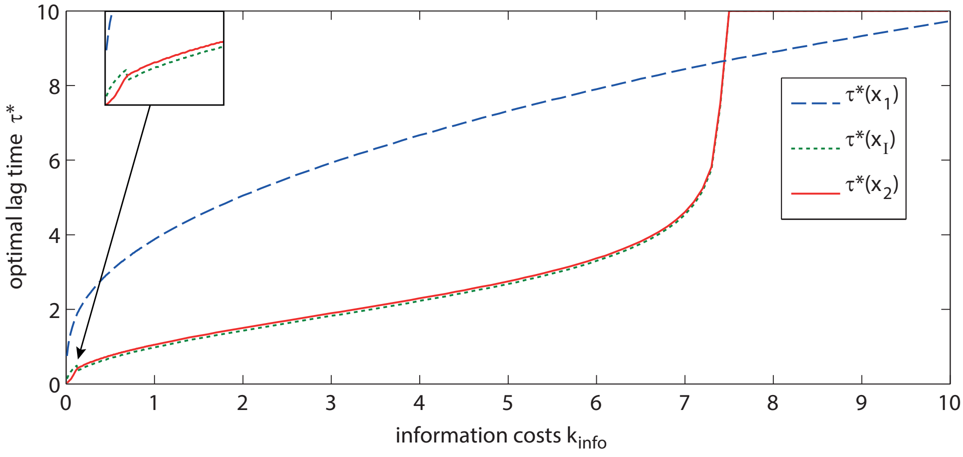

Figure 2.18: **$k_{\text{info}}$ vs. optimal lag times for the 3-state-example.** The optimal lag times $\tau^*(x)$ for Example 2.21 depending on the information costs $k_{\text{info}}$. At $k_{\text{info}} = 0.12$ the optimal lag time $\tau^*(x_I)$ of the intermediate state performs a jump.

Again, one could formulate "conditional" optimal policies, fixing the action for the intermediate state $x_I$, which would lead to a continuity and monotonicity of the lag times depending on $k_{\text{info}}$, compare Example 2.21.
For vanishing information costs the optimal lag times seem to converge to zero, which gives a reasonable connection to the original Markov control model. To verify this property we have to assume that state testing is generally profitable, which will be done in the following theorem.

**Theorem 2.37** (Continuity of $\tau^*$ with respect to $k_{\text{info}}$ at $k_{\text{info}} = 0$)**.** *Assume the action space $\mathcal{A}$ to be finite. Let $\mathrm{a}_0$ denote the optimal policy (i.e. optimal choice of action) of the original control problem (without information costs), and let $\mu_0$ be the corresponding equilibrium distribution. Let $x \in \mathcal{S}$ be such that $\mu_0(x) > 0$ and assume that there exists a state $y \in \mathcal{S}$ with $\mathrm{a}_0(x) \neq \mathrm{a}_0(y)$ and $\mathbb{P}_{\mathrm{a}_0(x)}(X_t = y|X_0 = x) > 0$ for some $t > 0$.*
*Then it holds*

$$\tau^*_{k_{\text{info}}}(x) \overset{k_{\text{info}}\to 0}{\longrightarrow} 0,$$

*where $\tau^*_{k_{\text{info}}}(x)$ is the optimal lag time for state $x$ given the information costs $k_{\text{info}}$.*

*Proof.* The justification is analogous to the proof of Theorem 2.22: Assuming that $\tau^*(x)$ does not converge to zero means that there exist periods in time where the process is not optimally controlled, compare Figure 2.12. This leads to a difference in the costs for each time the process is in state $x$. As $\mu_0(x) > 0$, this difference has an impact on the long-term average costs, as well. This means that the optimal long-term average costs $\eta^*_{\text{info}}$ of the information cost model would not converge to the optimal long-term average costs $\eta_0$ of the original problem, which is a contraction to Theorem 2.36. □

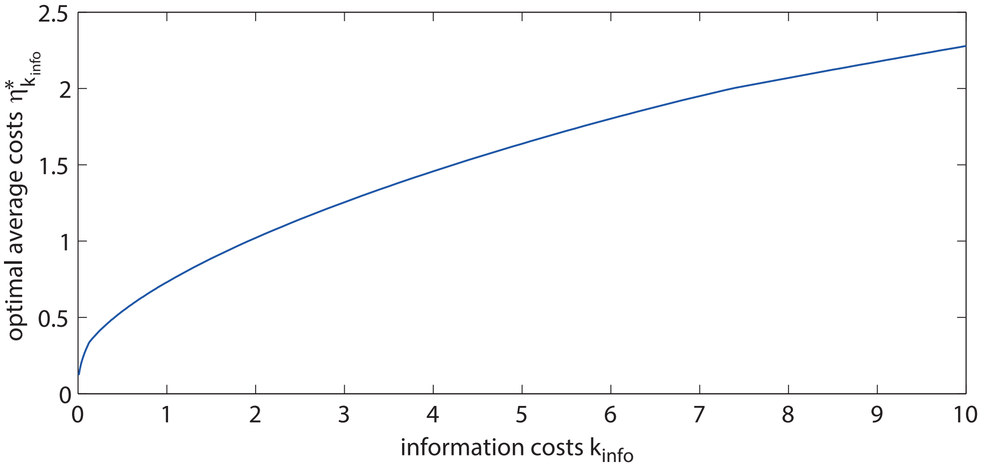

Figure 2.19: $\boldsymbol{k}_{\mathbf{info}}$ **vs. optimal average costs for the 3-state-example.** The constant of optimal average costs for Example 2.21 depending on the information costs $k_{\mathrm{info}}$, compare Theorem 2.36.

### 2.3.4 Sensitivity with respect to Lag Times $\tau(x)$

We end this section by presenting a short sensitivity analysis with respect to the lag times $\tau$ in the case of average costs. The motivation remains the same: Due to practical restrictions, an exact adherence of the calculated optimal lag times might not be possible in a real application. In the following, we question the possible consequences of such an irregularity.

According to the analysis of Section 2.3.1, we know that both the generators $G_u$ and the cost function $\tilde{C}_u$ are continuous with respect to $\tau(x)$ for all $x \in \mathcal{S}$ which implies that the average costs $\eta_u$ are continuous with respect to the lag time parameter. This is a reassuring fact in the sense that deviations from the optimal lag times – as long as they are not too large – will not lead to crucial changes in the costs.
In analogy to Section 2.2.4, we would like to see how $\eta_u$ depends on $\tau$ for the 2-state-example 2.25. Figure 2.20 shows the impact of changes in $\tau(x_1)$ resp. $\tau(x_2)$ when all other parameters are fixed to be optimal. In both cases, the net costs $\eta_{\mathrm{net}}$ are monotonously increasing in $\tau$, whereas the total costs $\eta$ exhibit a unique minimum. However, around these minima, there exists a wide area of values where the constant $\eta$ of long-term average costs is nearly constant, which indicates a low sensitivity with respect to the lag times for this concrete example.

As for the numerical consequences of such a low sensitivity we can argue as in the case of discounted costs, compare Section 2.2.4. On the one hand, an exact minimization by numerical methods would require an enormous number of iteration steps. On the other hand, the exact adherence of a prescribed lag time might not be realistic anyway, such that a nearby optimal solution will be satisfying.

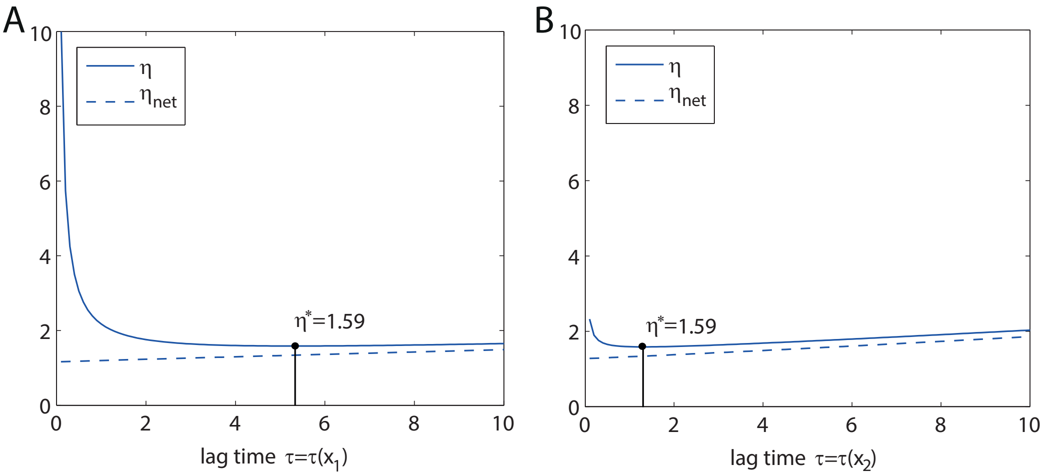

Figure 2.20: **Sensitivity with respect to $\tau$ for the 2-state-example.**
A: Lag time $\tau(x_1)$ vs. long-term average costs $\eta$ and $\eta_{\text{net}}$ for fixed $\tau(x_2) = 1.3$ and $a(x_1) = 1$, $a(x_2) = 2$. B: Lag time $\tau(x_2)$ vs. long-term average costs $\eta$ and $\eta_{\text{net}}$ for fixed $\tau(x_1) = 5.3$ and $a(x_1) = 1$, $a(x_2) = 2$.
All other parameters coincide with those given in the first line of Table 2.2, i.e. $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_2) = 2$, $k_{\text{info}} = 1$. The minimum of the total average costs $\eta$ is attained at $\tau^*(x_1) = 5.3$ (panel A) resp. $\tau^*(x_2) = 1.3$ (panel B) with $\eta = \eta^* = 1.59$ which is consistent with the values of the optimal policy declared in Table 2.2.

## 2.4 Comparing Discounted and Average Costs

In some situations it might not be clear which of the two presented cost criteria is more appropriate for a given application. In this regard, it is interesting to analyze how these criteria are related to each other: Given a fixed control problem, how do the optimal policies differ from each other? How are the optimal costs related to each other? Is there a case in which both criteria lead to the same optimal policy? These questions will be answered in the following.

Looking once again at the 2-state-process in Example 2.11 and Example 2.25 with $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_2) = 2$ and $k_{\text{info}} = 1$ (first line of Table 2.1 and Table 2.2) one can observe that the choice of actions coincides for both criteria, while time intervals and value function are different. Table 2.4 shows the optimal policies and the value functions for the discounted-cost criterion and different discount factors $\lambda$.

We can observe that, with decreasing $\lambda$, the values $\tau^*(x)$ of optimal lag times of the discounted-cost problem decrease, while the value function increases. The (relative) difference between the values $V_\lambda(x_1)$ and $V_\lambda(x_2)$ tends to zero. For $\lambda = 10^{-4}$, the lag times $\tau^*(x_1)$ and $\tau^*(x_2)$ approximately coincide with those of the average-cost criterion, and for the value function it holds

$$V_\lambda(x_1) \approx V_\lambda(x_2) \approx \frac{1}{\lambda}\eta^*.$$

| $\lambda$ | $a^*(x_1)$ | $a^*(x_2)$ | $\tau^*(x_1)$ | $\tau^*(x_2)$ | $V_\lambda(x_1)$ | $V_\lambda(x_2)$ |
|---|---|---|---|---|---|---|
| $10^{-1}$ | $a_1$ | $a_2$ | 11.3 | 1.8 | 7.8 | 69.8 |
| $10^{-2}$ | $a_1$ | $a_2$ | 5.7 | 1.4 | 145.2 | 255.1 |
| $10^{-3}$ | $a_1$ | $a_2$ | 5.4 | 1.3 | $1.57 \cdot 10^3$ | $1.69 \cdot 10^3$ |
| $10^{-4}$ | $a_1$ | $a_2$ | 5.3 | 1.3 | $1.59 \cdot 10^4$ | $1.59 \cdot 10^4$ |

Table 2.4: **Optimal policy and value function of discounted costs depending on $\boldsymbol{\lambda}$.** This table shows the dependence of the optimal policy and the value function on the discount factor $\lambda$ for the 2-state-example 2.11 and fixed cost parameters $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_2) = 2$ and $k_{\text{info}} = 1$. The corresponding results for the average-cost criterion are: $a^*(x_1) = a_1$, $a^*(x_2) = a_2$, $\tau^*(x_1) = 5.3$, $\tau^*(x_2) = 1.3$, $\eta^* = 1.59$, see Table 2.2.

Why are the values $\tau^*(x)$ larger for the discounted-cost criterion than for the average-cost criterion? In other words, why does the optimal policy for discounted costs afford a smaller number of tests? The reason should be the following: The damage which arises from rare testing (because of a delayed/suboptimal switch of action choice) is less valued because it is discounted!

The relationship

$$\eta_u = \lim_{\lambda \searrow 0} \lambda J_\lambda(x, u)$$

is well known for the original model [26]. However, we only translated the average-cost problem to an equivalent original Markov control model, while for the discounted-cost criterion a direct approach was feasible. The observation $\eta_u = \lim_{\lambda \searrow 0} \lambda J_\lambda(x, u)$ for the information cost model thus suggests that for small $\lambda$ also the discounted-cost problem is related to the original Markov decision process $(Y_t)_{t \geq 0}$ with generator $G_u$ and cost function $\tilde{C}$ which was derived from the average-cost problem, compare (2.33) and (2.27).

Chapter 3

# Application to Treatment Scheduling in HIV-1

Controlling a random process that cannot be observed without effort is a realistic situation. It is common practice to pay for information – whether by money or indirectly by spending time or energy. A central aspect is to decide when to gather the information in order to permit a satisfying control without producing too many information costs. Especially situations in a medical context follow this scheme; a medical practitioner does not only have to decide which medicine to prescribe, but must also find a date for the next checkup. His decision will depend on the actual health status of the patient and the dynamics of the disease process given the medical treatment. The Markov control model which was developed in Chapter 2 exactly reflects this interplay of information purchase and control adaption. The essential finding was the reformulation of the Bellman equation which provides a numerical scheme for the calculation of the optimal policy by means of dynamic programming. The results will now be applied to a detailed example in the medical context. We will consider the dynamics of the human immunodeficiency virus (HIV) and determine – from the national economic perspective – optimal therapeutic policies in resource-rich and resource-poor settings.

Optimal control methods have previously been applied by other research groups in the context of HIV-therapy: Luo et al. [44] and Vargas et al. [29] treated the underlying system deterministically, which fails to capture the intrinsic stochastic nature of HIV drug resistance development and the time-scales on which drug resistance develops [52]. Furthermore, it does not allow for individualized (patient-specific) treatment optimization. Shechter et al. [61] used MDPs (machine maintenance approaches) to maximize expected residual lifetime under treatment which allows for patient-specific treatment optimization but does not take into account the "cost of observation" and neither the virological state in a patient. Both aspects, however, will play an essential role and will be thoroughly addressed in our analysis.

In Section 3.1 we will define the parameters of the Markov control model

$$\Big(\mathcal{S},\, \mathcal{A},\, \{\mathcal{A}(x) : x \in \mathcal{S}\},\, \{L_a : a \in \mathcal{A}\},\, c\,,\, k_{\text{info}}\Big)$$

for the given application. In this regard, the most critical step is the definition of the cost function. Here, we have to assign costs to health damage and death – but does this not reflect a somehow unconscionable intend? Under ethical consideration, the death of a person cannot be measured in any form of monetary currency, as the life of an individual is priceless. However, medical treatment does produce a significant amount of costs and especially resource-poor countries cannot afford an arbitrarily individualized therapy which does not underlie any monetary restrictions. Instead, from the economic point of view the arrangement of the health system in fact requires a weighting of costs and benefit. We will handle the situation by taking the gross domestic product (GDP) of a country as a reference and measuring the costs of health damage in terms of the productivity loss with respect to the GDP.

Given the model for the HIV dynamics, we will in Section 3.2 determine the optimal policy for a resource rich and a resource poor setting. We will further calculate the cost splitting for different values of the information cost parameter $k_{\text{info}}$ and compare the optimal policy with the two extreme cases of continuous control (in the sense of original Markov control theory) and constant control (referring to fixing the action for all times).

In Section 3.3 we will analyze how deviations from the optimal lag times influence the process costs. Finally, the impact of cost reductions on the survival benefit will be considered in Section 3.4.

## 3.1 HIV Dynamics Model

In the following, we will introduce the HIV-model to which we will apply the theory of Markov control with information costs developed in Chapter 2. First of all, the state space $\mathcal{S}$ and the action space $\mathcal{A}$ of the HIV-model will be defined. Then we will explain how distinct treatments $a \in \mathcal{A}$ manipulate the entries of the generators $L_a$ and parametrize the corresponding cost function.

### State space $\mathcal{S}$ and action space $\mathcal{A}$

HIV dynamics and drug resistance development can accurately be described by stochastic reaction kinetics [52, 68, 69]. The fundamental evolution equation for stochastic kinetics is the chemical master equation (CME), for which each state comprises a combination of discrete numbers of individuals of the respective species (e.g. viral strains), resulting in state space dimensions $\mathbb{N}_0 \times \mathbb{N}_0 \times ... \times \mathbb{N}_0$, which is numerically infeasible in terms of a direct solution.

In order to reduce the dimensionality of the state space, we consider four combined states of copy numbers for each virus type. If a virus type is absent, we denote the respective state by 0; if it is present in low copy numbers, i.e. for $< 50$ virus copies/mL blood (detection limit of assays used in the clinic), the respective state is denoted by $\ell$; for medium copy numbers between 50 and 4000 virus copies/mL blood we denote the state by $m$; and for high copy numbers with more than 4000 virus copies/mL blood the state is denoted by $h$. This coarse graining is in line with the levels of virus produced in the distinct cellular reservoirs of HIV, see e.g. [69]. The $\ell$-states are reflecting states, which is justified by inability to eradicate HIV (the persistence of virus in reservoirs [21, 40]), and the $h$-states are reflecting states because there is a maximum carrying capacity of the system (i.e. the virus does not grow indefinitely). Further, the $\ell$-states do not affect patient health (thus not producing state costs) as the virus is essentially suppressed [11]. Costs are produced by the $h$-states and the $m$-states respectively, and the $h$-states produce more costs than the $m$-states (denoted later in Table 3.2).

The dynamics of the virus may be influenced by different lines of medical treatment. In line with [77] and [78], we choose the set of actions

$$\mathcal{A} = \{a_\emptyset, a_1, a_2\},$$

where $a_\emptyset$ denotes the absence of medical intervention, while $a_1$ and $a_2$ denote the application of two distinct treatment lines. This choice is motivated by the fact that in the following we will focus on HIV treatment in resource-constrained settings in which only two treatment lines ($a_1$ and $a_2$) are available.

According to their treatment susceptibility, the model distinguishes 4 viral strains $M$ ("mutants"): a strain WT (wild type) that is susceptible to both treatment lines, a strain R1 which is susceptible to $a_2$ but unaffected by (resistant to) $a_1$, a strain R2 that is susceptible to $a_1$ but unaffected by $a_2$ and a highly resistant strain HR which is resistant to both treatments.
Considering all permutations of viral strains $M \in \{\mathrm{WT}, \mathrm{R1}, \mathrm{R2}, \mathrm{HR}\}$ and respective copy numbers $n_C(M) \in \{0, \ell, m, h\}$ as well as patient death ✠, the state space of the corresponding Markov control model turns out to be

$$\mathcal{S} = \{0, \ell, m, h\}^4 \cup ✠$$

with $|\mathcal{S}| = 4^4 + 1 = 257$ states in total.
In order to describe a state $x \in \mathcal{S}$, we choose a compact vector notation of the form

$$x = \big[\, n_C(\mathrm{WT})\,,\; n_C(\mathrm{R1})\,,\; n_C(\mathrm{R2})\,,\; n_C(\mathrm{HR}) \,\big].$$

For example, the state $x = \big[h, \ell, m, 0\big]$ describes the situation of a $h$igh number of wild type strains, a $\ell$ow number of R1-mutants, a $m$edium number of R2-mutants and the absence of highly resistant mutants. We use this notation also for sets of states by writing, e.g. $\big[\{m, h\}, *, 0, 0\big]$, which stands for a $m$edium **or** $h$igh number of wild type strains, an arbitrary number of R1-mutants and the absence of R1-mutants and highly resistant mutants.

### Generator entries

For each action $a \in \{a_\emptyset, a_1, a_2\}$ the generator $L_a$ is a matrix in $\mathbb{R}^{|\mathcal{S}|,|\mathcal{S}|}$ containing the transition rates between the different states given this action. For the considered HIV-model the transitions between states can be split up into three categories:

First, the copy number $n_C(M)$ can increase or decrease for each viral strain $M$. Second, a patient can die such that the process switches to state ✠. Third, there are transitions between the different viral strains due to mutations. By the structure of the HIV-model some of these rates naturally should be zero: For example, there is no direct switch from a low copy number to a high copy number, instead such a growth process always runs through the state of a medium copy number. All the non-zero entries of the generators will now be explained, starting with the first category.

The basic transitions between copy number states $n_C(M)$ for a particular viral strain $M$ (here exemplified for the wild type strain WT) of our continuous-time Markov model are the following:

$$\big[\ell\,,*\,,*\,,*\big] \underset{\delta_m}{\overset{k_{\ell,a}}{\rightleftarrows}} \big[m\,,*\,,*\,,*\big], \qquad \big[m\,,*\,,*\,,*\big] \underset{\delta_h}{\overset{k_{m,a}}{\rightleftarrows}} \big[h\,,*\,,*\,,*\big], \quad (3.1)$$

where $*$ indicates an arbitrary number of the respective virus strain (R1, R2 and HR in the example above). The parameters $k_{\ell,a}$ and $k_{m,a}$ denote the reaction rates for a transition from copy number $\ell$ to copy number $m$ and from copy number $m$ to copy number $h$, respectively (viral growth), which depend on the treatment $a \in \{a_\emptyset, a_1, a_2\}$. The parameters $\delta_m$ and $\delta_h$ are independent of the treatment and denote the reaction rates of going from copy number $m$ to copy number $\ell$ and from copy number $h$ to copy number $m$, respectively (virus elimination).

The occurrence of death is considered in the following way:

$$\big[h\,,*\,,*\,,*\big] \xrightarrow{d_h} \maltese, \quad \big[m\,,*\,,*\,,*\big] \xrightarrow{d_m} \maltese, \quad \big[\ell\,,*\,,*\,,*\big] \xrightarrow{d_\ell} \maltese. \qquad (3.2)$$

The parameters $d_h > d_m > d_\ell$ denote the rate for the death of the patient. These parameters are also unaffected by the treatments. We assume that high viral burden (states $h$ and $m$ respectively) increases the risk of death, whereas $d_\ell$ equals the rate for "natural death". The rate for natural death is computed according to $d_\ell = 1/(\text{residual life expectancy healthy})$ and is exemplified in the caption of Table 3.2. Likewise, $d_h$ and $d_m$ are computed using the average residual life expectancy in states $h$ and $m$.
A visualization of the transitions between copy numbers and the occurrence of death is given in Figure 3.1 (left).

The considered transitions (mutations) between viral strains $M$ are depicted in Figure 3.1 (right). Specifically, mutation generates a $\ell$ow number of viral particles from either a $m$edium or $h$igh number of viruses belonging to a distinct strain. Exemplified for the wild type strain WT those are:

$$\begin{aligned}
\big[h\,,0\,,*\,,*\big] &\xrightarrow{\mu_{h,\mathrm{R1},a}} \big[h\,,\ell\,,*\,,*\big], & \big[m\,,0\,,*\,,*\big] &\xrightarrow{\mu_{m,\mathrm{R1},a}} \big[m\,,\ell\,,*\,,*\big], & (3.3)\\
\big[h\,,*\,,0\,,*\big] &\xrightarrow{\mu_{h,\mathrm{R2},a}} \big[h\,,*\,,\ell\,,*\big], & \big[m\,,*\,,0\,,*\big] &\xrightarrow{\mu_{m,\mathrm{R2},a}} \big[m\,,*\,,\ell\,,*\big], & (3.4)\\
\big[0\,,h\,,*\,,*\big] &\xrightarrow{\mu_{h,\mathrm{R1},a}} \big[\ell\,,h\,,*\,,*\big], & \big[0\,,m\,,*\,,*\big] &\xrightarrow{\mu_{m,\mathrm{R1},a}} \big[\ell\,,m\,,*\,,*\big], & (3.5)\\
\big[0\,,*\,,h\,,*\big] &\xrightarrow{\mu_{h,\mathrm{R2},a}} \big[\ell\,,*\,,h\,,*\big], & \big[0\,,*\,,m\,,*\big] &\xrightarrow{\mu_{m,\mathrm{R1},a}} \big[\ell\,,*\,,m\,,*\big]. & (3.6)
\end{aligned}$$

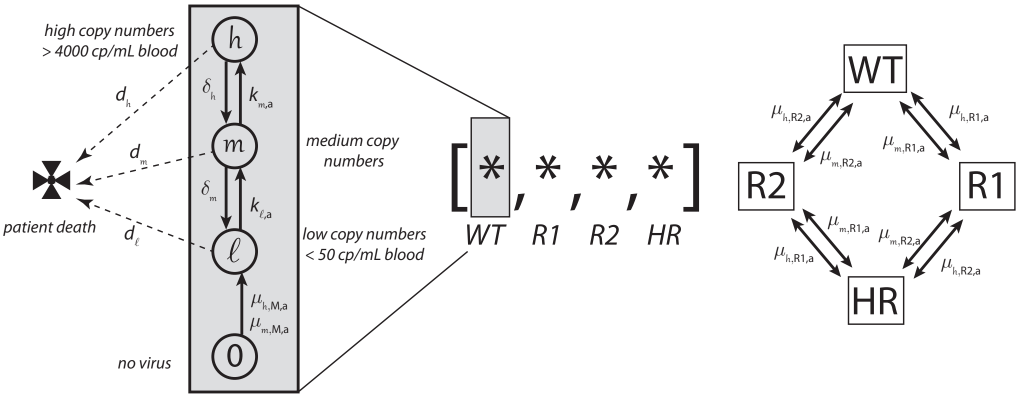

Figure 3.1: **Simplified HIV Model.** Left: Transitions between copy numbers $n_C \in \{0, \ell, m, h\}$ within a viral strain $M \in \{\mathrm{WT}, \mathrm{R1}, \mathrm{R2}, \mathrm{HR}\}$ and occurrence of death. Right: Transitions in between different viral strains $M$. Taken from [77].

The lines (3.3) and (3.4) indicate mutation arising from the wild type strain and the lines (3.5) and (3.6) indicate mutations yielding the wild type strain. The parameters $\mu_{h,\mathrm{R1},a}$ and $\mu_{h,\mathrm{R2},a}$ denote the propensity for the emergence or disappearance of a mutation that confers drug resistance to treatment $a_1$ and $a_2$, respectively, emanating from copy number state $h$. Analogously, $\mu_{m,\mathrm{R1},a}$ and $\mu_{m,\mathrm{R2},a}$ denote the propensity for the emergence or disappearance of a mutation emanating from copy number state $m$. Note that we consider only the following mutations: WT ↔ R1, WT ↔ R2, R1 ↔ HR and R2 ↔ HR, which is motivated by the fact that a direct transition from WT ↔ HR is very unlikely, because the genetic distance between the two viral strains is too large to be overcome at once.

The effect of the treatments $a_1$ and $a_2$ on the growth and mutation rates is considered in the following way:

$$k_{\ell,a} = \Big(1 - \eta(a,M)\Big) k_{\ell,\emptyset}, \qquad k_{m,a} = \Big(1 - \eta(a,M)\Big) k_{m,\emptyset}, \tag{3.7}$$

$$\mu_{h,\tilde{M},a} = \Big(1 - \eta(a,M)\Big) \mu_{h,\tilde{M},\emptyset}, \qquad \mu_{m,\tilde{M},a} = \Big(1 - \eta(a,M)\Big) \mu_{m,\tilde{M},\emptyset}, \tag{3.8}$$

where $M \in \{\mathrm{WT}, \mathrm{R1}, \mathrm{R2}, \mathrm{HR}\}$ is the present viral strain which induces the reaction.[6] The parameter $\eta(a, M)$ denotes the efficacy of treatment $a \in \{a_1, a_2\}$ on this viral strain $M \in \{\mathrm{WT}, \mathrm{R1}, \mathrm{R2}, \mathrm{HR}\}$; i.e. if strain $M$ is susceptible to treatment $a \in \{a_1, a_2\}$, then $0 < \eta(a, M) \leq 1$, and if the viral strain $M$ is insusceptible to treatment $a \in \{a_1, a_2\}$ then $\eta(a, M) = 0$. The parameters $k_{\ell,\emptyset}$ and $k_{m,\emptyset}$ resp. $\mu_{h,\tilde{M},\emptyset}$ and $\mu_{m,\tilde{M},\emptyset}$ denote the growth rates resp. mutation rates in the absence of intervention, i.e. for $a_\emptyset$ (see Table 3.1).

[6] In fact, this means that the described parameters additionally depend on the considered viral strain, i.e. one would have to write $k_{\ell,a}(M)$, $\mu_{h,\tilde{M},a}(M)$ in order to notice this dependency. Here we omit this additional parameter for the purpose of legibility.

We estimated all parameters by fitting the presented model to clinical data of virus decay- and rebound dynamics, choosing a least square criterion, see [77]. Some of the described rates are country specific, while others can be chosen generically. The generic model parameters are given in Table 3.1, while the country specific parameters are shown in Table 3.2.

| param. | value | param. | value | param. | value |
|---|---|---|---|---|---|
| $\delta_h$ | $6.13 \cdot 10^{-2}$ | $\mu_{h,\mathrm{R1},\emptyset}$ | 1.24 | $\eta(a_1, \{\mathrm{WT}, \mathrm{R2}\})$ | 0.979 |
| $\delta_m$ | $5.1 \cdot 10^{-2}$ | $\mu_{m,\mathrm{R1},\emptyset}$ | $4.34 \cdot 10^{-2}$ | $\eta(a_1, \{\mathrm{R1}, \mathrm{HR}\})$ | 0 |
| $k_{\ell,\emptyset}$ | 0.13 | $\mu_{h,\mathrm{R2},\emptyset}$ | $2.41 \cdot 10^{-4}$ | $\eta(a_2, \{\mathrm{WT}, \mathrm{R1}\})$ | 0.966 |
| $k_{m,\emptyset}$ | 0.13 | $\mu_{m,\mathrm{R2},\emptyset}$ | $2.33 \cdot 10^{-2}$ | $\eta(a_2, \{\mathrm{R2}, \mathrm{HR}\})$ | 0 |

Table 3.1: **Generic model parameters.** All parameters in units [1/day] except $\eta$ [unit less]. For the death rates $d_h$, $d_m$, $d_\ell$ see Table 3.2.

### Cost parameters

Our analysis is conducted from a country's or a public health-care/monetary perspective. The costs are produced by the health damage of a patient on the one hand (giving the state costs) and by the price for medical intervention on the other hand (giving the action costs), such that the cost function is of the form $c(x, a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$. The costs $c_{\mathcal{S}}(x)$ of being in the respective states $x \in \mathcal{S}$ are computed based on the average productivity loss $\mathrm{pL}(n_C)$ (depending on the copy number of a viral strain) times the average daily monetary contribution of one individual (assessed in terms of daily per capita GDP), i.e.

$$c_{\mathcal{S}}(x) = \mathrm{pL}(x) \cdot \mathrm{GDP}$$

with $\mathrm{pL}(x) := \max_M \; \mathrm{pL}(n_C(M))$ for $x \in \{0, \ell, m, h\}^4$, and $\mathrm{pL}(\text{✠}) = 1$, which means that death is interpreted in terms of a complete loss in productivity. The values are adapted from [58].

We tend to assess optimal policies in the case where two treatment lines ($a_1$ and $a_2$) are available in (i) developed countries with Germany as a representative, and (ii) in resource-constrained settings exemplified for South Africa because of the extraordinary high prevalence (17.8%) of HIV in this country [67]. The daily costs $c_{\mathcal{A}}$ of treatment in Germany and in resource-constrained settings were derived from [64, 65]. In resource-constrained settings, the William J. Clinton Foundation has negotiated prices for antiviral drugs which are highly subsidized, giving access to antivirals in these settings. The respective parameters are displayed in Table 3.2. The costs for drug resistance testing are $k_{\text{info}} \approx 400$ US$ per test in the western world [58]. In resource-constrained settings, these tests are not subsidized. Furthermore, because of the often undeveloped infrastructure tests may be even more expensive, thus justifying the adjustment to a steeper value of $k_{\text{info}} = 500$ US$ per test. All parameters related to costs are displayed in Table 3.2.

| param. | value | value | unit | ref. |
|---|---|---|---|---|
| | **Germany** | **S. Africa** | | |
| $c_{\mathcal{A}}(a_{\emptyset})$ | 0 | 0 | - | - |
| $c_{\mathcal{A}}(a_1)$ | 48.5 | 0.3 | US$/d | [64, 65] |
| $c_{\mathcal{A}}(a_2)$ | 58.8 | 1.08 | US$/d | [64, 65] |
| $k_{\text{info}}$ | 400 | 500 | US$/d | [58] |
| $d_\ell$ | $6.2 \cdot 10^{-5}$ | $9.4 \cdot 10^{-5}$ | 1/d | ♮ |
| $d_m$ | $2.7 \cdot 10^{-4}$ | $2.7 \cdot 10^{-4}$ | 1/d | † |
| $d_h$ | $5.5 \cdot 10^{-4}$ | $5.5 \cdot 10^{-4}$ | 1/d | † |
| GDP | 43 742 | 8 066 | US$/p.p./y | [66] |
| pL($\ell$) | 0 | 0 | - | [58] |
| pL($m$) | 0.1 | 0.1 | - | [58] |
| pL($h$) | 0.4 | 0.4 | - | [58] |
| pL(✠) | 1 | 1 | - | - |
| $\lambda$ | $1 \cdot 10^{-4}$ | $1.75 \cdot 10^{-4}$ | 1/d | ‡ |

Table 3.2: **Country specific model parameters.** Here, $k_{\text{info}}$ refers to the price for a drug resistance test and $\lambda$ is the discount factor. The GDP refers to the estimation for the year 2011 by the International Monetary Fund. ♮ Computed from the overall residual life expectancy normalized by AIDS related death. The overall life expectancy in Germany and South Africa is 79.4 and 49.3 years, respectively, with an average age of HIV detection of 35 and 25 years and HIV prevalence of 0.1% and 17.8%, respectively. † For states $m$ and $h$ we assumed a respective residual life expectancy of 10 and 5 years. ‡ Assuming an annual inflation of 3.5% and 6.2% for Germany and South Africa, respectively.

**Remark 3.1.** *The presented HIV-model is based on the aggregation of possible copy numbers into four combined states which strongly coarsens the true dynamics. Of course, a consideration of the exact master equation would be a more accurate approach. However, such a detailed approach would cause two problems: First, the parameters of the exact model are difficult to determine, and second, the detailed model would enormously extend the run time of the considered optimization algorithms due to its giant state space. Finding a way to handle these problems clearly exceeds the scope of this thesis.*

*It is common practice – especially in the context of molecular dynamics – to approximate complicated dynamical behaviour by low dimensional models in order to reduce numerical effort, see e.g. [12, 13, 16, 18, 53, 57]. Typically, the essential dynamics of a system are detected by identifying almost invariant (or metastable) sets of the state space. The combined states of copy numbers (0, $\ell$, $m$, $h$) of the presented HIV-model are metastable in the sense that the process of virus growth typically stays in each of these states for a comparably long time.*

*For the real-world application another point is relevant: Using the detailed model would require to determine the exact copy number of the virus in the body of a patient by medical tests which is not realistic.*

## 3.2 Optimal Control and Cost Splitting

In the following, we will apply the developed Markov control theory of Chapter 2 to treatment scheduling and diagnostic testing in HIV-1 using the model presented in Section 3.1. As the model contains an absorbing state (the death of a patient), it only makes sense to consider the discounted-cost criterion defined in (2.3). The discount factor $\lambda$ is chosen with reference to the inflation in the considered countries, see Table 3.2.

The optimal control is determined by the discounted-cost policy iteration algorithm presented on page 55. The results are displayed in Table 3.3, both for the resource-rich and the resource-poor setting, given the cost parameters of Table 3.2. The composition of the value function for state $x = \big[h\,,\,0\,,\,0\,,\,0\big]$ and its dependence on the cost parameters is analyzed subsequently.

| states | $a^*(x)$ | $\tau^*(x)$ | $a^*(x)$ | $\tau^*(x)$ |
|---|---|---|---|---|
| | **Germany** | | **South Africa** | |
| $\big[\ell\,,\,0\,,\,0\,,\,0\big]$ | $a_1$ | 155 | $a_1$ | $\geq \tau_{\max}$ |
| $\big[\{m, h\}\,,\,0\,,\,0\,,\,0\big]$ | $a_1$ | $6 - 24$ | $a_1$ | $11 - 45$ |
| $\big[*\,,\,0\,,\,\{\ell, m, h\}\,,\,0\big]$ | $a_1$ | $20 - 554$ | $a_1$ | $\geq \tau_{\max}$ |
| $\big[*\,,\,\{\ell, m, h\}\,,\,0\,,\,0\big]$ | $a_2$ | $159 - 567$ | $a_2$ | $\geq \tau_{\max}$ |
| otherwise | $a_\emptyset$ | $\geq \tau_{\max}$ | $a_1$ | $\geq \tau_{\max}$ |

Table 3.3: **Optimal policy.** Calculated optimal policy for the resource-rich (Germany) and resource-poor setting (South Africa) giving the optimal choice of treatment $a^*(x)$ and the optimal lag time $\tau^*(x)$ (in days) depending on the state of the patient. For clarity reasons, states are merged according to their related treatment choice. The values given for $\tau^*$ refer to the respective minimum and maximum value of $\tau^*(x)$ for the states $x$ indicated in the first column; e.g. for the second row it holds $\tau^*(\big[h, 0, 0, 0\big]) = 6$ and $\tau^*(\big[m, 0, 0, 0\big]) = 24$ for Germany. For the numerical computations we chose $\tau_{\max} = 2\,000$ days.

According to Table 3.3, treatment $a_1$ is chosen whenever only wild type (WT) virus is present or wild type (WT) and strains resistant to treatment $a_2$ (R2) coincide. Treatment $a_2$ is only chosen if drug resistance to treatment $a_1$ has emerged (R1) while the virus is still susceptible to treatment $a_2$. Interestingly, there is a difference in the handling of the other states (i.e. highly resistant strains HR, or the concurrence of R1 and R2): While when considering Germany no treatment ($a_\emptyset$) is given, treatment $a_1$ is applied in the resource-poor setting. This result is due to the fact that the use of treatment in patients that carry drug resistant viruses may provide limited benefit in comparison to the treatment costs for Germany, whereas costs for treatment in resource-constrained settings are in fact so low that their application in the case of drug-resistant virus is still cost-optimal. This is also supported by the cost-splitting in Table 3.4 (baseline parameters in first column): For

Germany, treatment cost $c_{\mathcal{A}}$ produce more than 20% of the total optimal costs, whereas they only produce about 2.5% of the total costs in South Africa. In fact, STOLL et al. [64] argued that treatment may be too expensive in Germany because of the use of original manufacturer's drugs instead of generic drugs.

| | | $k_{\text{info}}$ | | | | |
|---|---|---|---|---|---|---|
| | | basel.$^{\star}$ | 200 | 100 | 50 | 5 |
| Germany | $V_{\text{info}}(x)$ | 10 502 | 7 681 | 5 393 | 3 684 | 950 |
| | $V_{\mathcal{A}}(x)$ | 208 100 | 211 150 | 212 950 | 214 030 | 215 510 |
| | $V_{\mathcal{S}}(x)$ | 705 380 | 698 180 | 693 710 | 690 900 | 686 880 |
| | $V_{\lambda}(x)$ | 923 982 | 917 011 | 912 053 | 908 624 | 903 340 |
| S. Africa | $V_{\text{info}}(x)$ | 2 294 | 2 467 | 1 838 | 1 314 | 369 |
| | $V_{\mathcal{A}}(x)$ | 1 739 | 1 899 | 1 965 | 1 909 | 1 842 |
| | $V_{\mathcal{S}}(x)$ | 65 116 | 62 408 | 60 928 | 60 928 | 60 104 |
| | $V_{\lambda}(x)$ | 69 149 | 66 774 | 64 731 | 64 151 | 62 315 |

Table 3.4: **Cost splitting.** Calculated cost splitting $V_{\lambda}(x) = V_{\mathcal{S}}(x) + V_{\mathcal{A}}(x) + V_{\text{info}}(x)$ (in US$) for state $x = \big[h, 0, 0, 0\big]$ in the resource-rich and resource-poor setting, respectively, using the results of Section 2.2.2. $^{\star}$ Baseline costs for resistance tests are 400 US$ and 500 US$ in Germany and South Africa, respectively.

As can be seen in Table 3.3, much longer periods between tests are proposed in the resource-constrained setting in comparison to the resource-rich setting. In fact, drug-resistance testing (and thus the ability to adapt one's individual therapy) is only recommended in states $\big[\{m, h\}, 0, 0, 0\big]$ in the resource-poor setting. It may therefore be indicated for the resource-constrained setting, that despite the availability of subsidized treatment, their optimal use may not be feasible because informed decision making is not possible as a consequence of unaffordable diagnostics ($k_{\text{info}}$ is too high). Note that drug resistance tests are currently not part of the standard of care in resource-constrained settings. In the resource-rich setting (Germany) information costs produce 1.1% of the total costs, whereas they produce 3.3% of total costs in the resource-poor setting (South Africa), see Table 3.4. Interestingly, the cost of information ($V_{\text{info}}$ in Table 3.4) is not reduced for South Africa when the price for diagnostics is reduced from $k_{\text{info}} = 500$ to $k_{\text{info}} = 200$ US$, while at the same time the total costs are reduced, which can be fully attributed to a state cost reduction. This indicates that the price reduction for diagnostics enables their more frequent use (thus no $V_{\text{info}}$ reduction), which seems to fully benefit the patient (lower $V_{\mathcal{S}}$). In total, state costs $V_{\mathcal{S}}$ are reduced by 2.6% in Germany when comparing the baseline parameters with $k_{\text{info}} = 5$ US$ per test, whereas they are reduced by 7.7% in South Africa. Finally, it can be seen that the total expected costs (last rows in Table 3.4) are disproportionately higher in Germany than in South Africa (compare their differences with the differences in GDP in Table 3.2).

**Example 3.2** (Trajectory under optimal control)**.** *A numerical simulation of the process under optimal control for the resource-poor setting delivered the trajectory displayed in Figure 3.2. The parameters of the optimal policy are taken from Table 3.3. After starting in $x_0 = [h, 0, 0, 0]$ at time $t_0 = 0$, action $a_1$ is applied and the process quickly switches to $[m, 0, 0, 0]$ which means a virus elimination. This is observed at time $t_1 = t_0 + \tau^*(x_0) = 11$ and leads to a further application of action $a_1$. The copy number of the wild type is again reduced, carrying the process to state $[\ell, 0, 0, 0]$; however, at the next observation time $t_2 = t_1 + \tau^*([m, 0, 0, 0]) = 56$ the process already returned to $[m, 0, 0, 0]$. Within the following time period of hidden progress, a mutation originates the viral strain $R1$ which successively grows. A test at time $t_3 = t_2 + \tau^*([m, 0, 0, 0]) = 101$ indicates this behaviour, and the action is adapted to $a^*([m, h, 0, 0]) = a_2$. From now on, by $\tau^*([m, h, 0, 0]) = \infty$, the action is fixed and no further tests are made.*

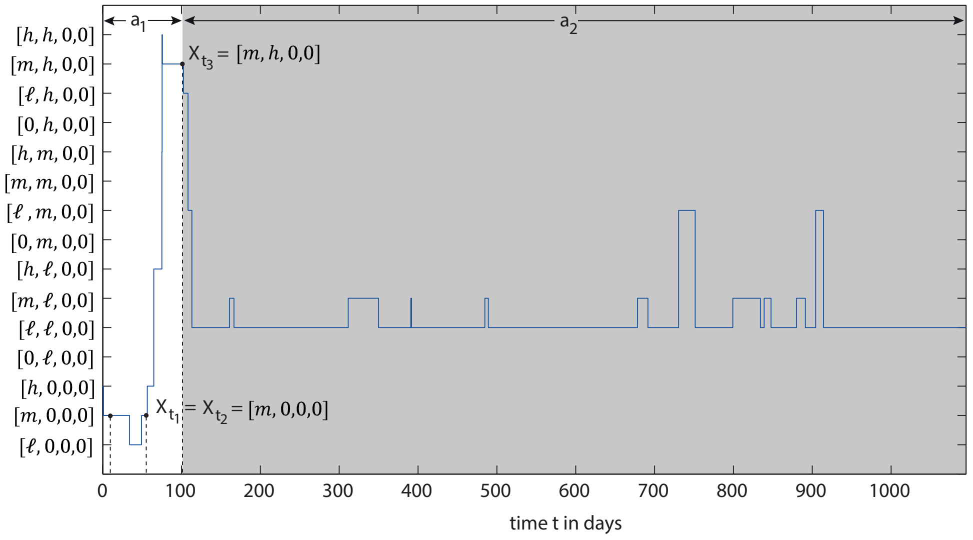

Figure 3.2: **Sample path.** Sample of virus dynamics given the optimal policy for South Africa, compare table 3.3. The gray area marks the time where action $a_2$ is applied, otherwise action $a_1$ determines the dynamics. The dotted lines indicate the points $t_n$ in time where tests are made and the state is observed.

### Comparison to constant control and original Markov control problem

As in the theoretical analysis of Chapter 2, the cost splitting provides a basis for a comparison to the original Markov control problem where the process is assumed to be completely observable. Such a comparison will be done in the following, both in terms of the net costs and in terms of the probability of death which is of special interest in the given application. As a contrasting extreme case, we consider the process under constant control, i.e. the action is fixed for all times. Comparing the optimal policy given in Table 3.3 to these two settings of (i) continuous control and (ii) constant control, we get an overview of the impact that different policies have on the process evolution.

(i) For the original Markov control setting we consider the discounted-cost criterion given in (1.2). The calculation of the optimal policy is conducted according to the discounted-cost policy iteration algorithm introduced in Section 1.2. The result is quite complex: Parallel to the optimal policies given in Table 3.3, both for Germany and for South Africa action $a_1$ is chosen whenever only the wild type (WT) is present or a combination of WT and strain R2 appears. If the wild type is accompanied by strain R1 (resistance to $a_1$), generally $a_2$ is the right choice, except for state $[m, h, 0, 0]$ which requires $a_1$. The reason seems to be the slightly higher efficiency $\eta(a_1, \mathrm{WT})$ of treatment $a_1$ with respect to the wild type: While the high copy number of viral strain R1 makes its control redundant (because its copy number can only decrease and the decrease rate is not effected by the actions), the wild type can still be prevented from growing further which justifies to apply the more effective treatment $a_1$.[7]
Such a pattern reappears in the resource-constrained setting (with interchanged roles of $a_1$ and $a_2$) when additionally R2 is present, i.e. for states of the form $[*, *, \{\ell, m, h\}, 0]$. In this case, action $a_1$ is optimal, except for those states with $n_C(R1) = m$ which require $a_2$. This time the reason is not only given by the growth rate of R1 (which is unaffected by $a_1$), but also by the mutation rates $\mu_{h,\mathrm{R2},\emptyset}$ and $\mu_{m,\mathrm{R2},\emptyset}$: Given $n_C(\mathrm{R1}) = m$, the value $\mu_{m,\mathrm{R2},\emptyset}$ ($> \mu_{h,\mathrm{R2},\emptyset}$) is reduced by $a_2$ making a mutation R1 → HR less probable. For $n_C(\mathrm{R1}) = h$, however, such a reduction is not required because $\mu_{h,\mathrm{R2},\emptyset}$ is very small already, compare Table 3.1, and so $a_1$ is the right choice.[8] In the presence of highly resistant strains (HR), both treatments $a_1$ and $a_2$ appear as optimal (depending on the state), and especially for the case of high copy numbers no medical intervention ($a_\emptyset$) is recommended, which is again due to the fact that the decrease rates are not effected by the (costly) treatments $a_1$ and $a_2$.
For the resource-rich setting, the structure is similar; however, there are much more states where $a_\emptyset$ has to be chosen, which can be explained by the high action costs in this setting.

(ii) The process under constant control refers to the situation in which an action is initially chosen and maintained for the remaining time. It can be associated to infinitesimal large information costs ($k_{\text{info}} = \infty$) which make testing unaffordable ($\tau(x) = \infty$ for all $x \in \mathcal{S}$). In the presented model this situation refers to the choice of either $a_\emptyset$, $a_1$, or $a_2$ for all times. The corresponding costs are given by

$$J_\lambda^a(x) := \mathbb{E}_x^a \left( \int_0^\infty e^{-\lambda s} c(X_s, a)\, ds \right) \tag{3.9}$$

with $a \in \{a_\emptyset, a_1, a_2\}$. Especially, the choice of $a = a_\emptyset$ stands for the "natural" disease

[7] Setting $\eta(a_1, \mathrm{WT}) = \eta(a_2, \mathrm{WT})$ yields $a_2$ for all states $[*, \{\ell, m, h\}, 0, 0]$.

[8] Setting $\mu_{m,\mathrm{R2},\emptyset} = \mu_{h,\mathrm{R2},\emptyset}$ avoids the switch to $a_2$ as long as $n_C(\mathrm{R2}) > \ell$.

| | $k_{\text{info}} = \infty$<br>constant control | | | $k_{\text{info}} = 500$<br>MDP wt.<br>info costs | $k_{\text{info}} = 0$<br>original<br>MDP |
|---|---|---|---|---|---|
| | $a_\emptyset$ | $a_1$ | $a_2$ | | |
| total costs | 107 350 | 76 790 | 70 030 | 69 149 | 61 420 |
| net costs | 107 350 | 76 790 | 70 030 | 66 855 | 61 420 |
| state costs | 107 350 | 76 024 | 66 940 | 65 116 | 59 589 |
| $\mathbb{P}_{x_0}(X_{3y} = \maltese)$ | 0.44 | 0.22 | 0.15 | 0.15 | 0.13 |
| $\mathbb{P}_{x_0}(X_{5y} = \maltese)$ | 0.63 | 0.34 | 0.24 | 0.23 | 0.21 |
| $\mathbb{P}_{x_0}(X_{15y} = \maltese)$ | 0.95 | 0.69 | 0.62 | 0.60 | 0.54 |

Table 3.5: **Comparison to constant control and original MDP for South Africa.** The net costs are given by $J_\lambda^a([h,0,0,0])$ in the case of constant control, by $V_\mathcal{S}([h,0,0,0]) + V_\mathcal{A}([h,0,0,0])$ in the case of MDP with information costs and by $\hat{V}_\lambda([h,0,0,0])$ in the case of original MDP. The probability of death $\mathbb{P}_{x_0}(X_t = \maltese) = \mathbb{P}(X_t = \maltese | X_0 = x_0)$ after 3, 5 or 15 years when starting in state $x_0 = [h,0,0,0]$ was computed by analytically solving the Kolmogorov equations in the case of constant control and in the original MDP setting, where we used the generator under optimal control $L^*(x,y) = L_{a^*(x)}(x,y)$. In the MDP with information costs setting, we approximated $\mathbb{P}(X_t = \maltese | X_0 = [h,0,0,0])$ using a well-established Monte-Carlo-Method [22].

process without medical intervention.

Naturally, it holds that

$$J_\lambda^a(x) \geq V_\lambda(x) \geq \hat{V}_\lambda(x) \quad \forall a \in \mathcal{A}, x \in \mathcal{S}, \tag{3.10}$$

where $\hat{V}_\lambda$ is the value function of the original Markov control model. The first inequality follows from the fact that the policy of constant control is contained in the set of policies $\mathcal{U}$ over which we minimize to determine $V_\lambda$, while the second inequality was motivated in Section 2.2.2. The same is true if we consider – instead of the total optimal costs $V_\lambda$ – the net costs $V_{\text{net}} = V_\mathcal{S} + V_\mathcal{A}$ which are the total costs without information costs. As both $J_\lambda^a$ and $\hat{V}_\lambda$ do not contain any information costs (in the original setting information is for free, and in the setting of constant control there are no tests at all) considering the net costs $V_{\text{net}}$ instead of $V_\lambda$ is better suited to make a comparison.

Table 3.5 shows the costs and the probability of death for the two settings (i) and (ii) as well as for $k_{\text{info}} = 500$ (compare Table 3.2) in the resource-constrained setting, choosing the initial state $x = [h,0,0,0]$. We can make the following observations. In accord with (3.10), the costs of the optimal MDP scheme with information costs go below those of any constant control, while they exceed the costs of the original MDP scheme with permanent optimal control. There is a huge difference between the costs resulting from the absence of medical treatment ($a_\emptyset$ at all times) and those arising under constant control with $a_1$ or $a_2$. While the values of the net costs all significantly differ from each other, the value of the total optimal costs

| | $k_{\text{info}} = \infty$<br>constant control | | | $k_{\text{info}} = 400$<br>MDP wt.<br>info costs | $k_{\text{info}} = 0$<br>original<br>MDP |
|---|---|---|---|---|---|
| | $a_\emptyset$ | $a_1$ | $a_2$ | | |
| total costs | 1 083 800 | 993 100 | 991 320 | 923 982 | 901 490 |
| net costs | 1 083 800 | 993 100 | 991 320 | 913 480 | 901 490 |
| state costs | 1 083 800 | 825 700 | 761 340 | 705 380 | 685 630 |
| $\mathbb{P}_{x_0}(X_{3y} = \maltese)$ | 0.44 | 0.20 | 0.12 | 0.11 | 0.11 |
| $\mathbb{P}_{x_0}(X_{5y} = \maltese)$ | 0.63 | 0.31 | 0.20 | 0.18 | 0.17 |
| $\mathbb{P}_{x_0}(X_{15y} = \maltese)$ | 0.95 | 0.64 | 0.57 | 0.49 | 0.47 |

Table 3.6: **Comparison to constant control and original MDP for Germany.** Again, the net costs are given by $J_a([h,0,0,0])$ in the case of constant control, by $V_{\mathcal{S}}([h,0,0,0]) + V_{\mathcal{A}}([h,0,0,0])$ in the case of MDP with information costs and by $\hat{V}([h,0,0,0])$ in the case of original MDP. The computational details agree with those described in Table 3.5.

$V([h,0,0,0]) = 69\,149$ arising from optimal control with information costs is very close to the costs under constant control with $a_2$. This means that in terms of the total costs, the optimal policy with information costs ($k_{\text{info}} = 500$) is only slightly better than a "blind"/constant control which could challenge the utility of medical testing. In fact, it is the reduction of action- and state costs which justifies the medical tests.

In terms of survival benefit, the optimal MDP scheme with information costs (information costs $k_{\text{info}} = 500$) is clearly better than the absence of medical intervention ($a_\emptyset$ at all times) and better than a constant treatment with only therapy line $a_1$, however, it is only slightly better than a constant treatment with $a_2$ for the time horizon analyzed (3, 5 and 15 years). For larger time horizons though, the differences between constant treatment with only $a_2$ and the optimal MDP scheme with information costs can be expected to further increase. Best in terms of survival benefit is the permanent optimal control of the original MDP model. The biggest difference in the probability of death can again be found when comparing the absence of medical intervention with all other considered policies. This emphasizes that the fundamental step is to start a medical treatment, while the details of the treatment policy are secondary.

The results for the resource-rich setting have the same structure, see Table 3.6, such that the analysis is completely analogous. It suffices to mention that the difference between the optimal control with information costs ($k_{\text{info}} = 400$) and the constant control with $a_2$ is more pronounced, whereas there is only a slight further improvement by permanent control. This highlights the positive effect that – even rare – state observations have on the process evolution as long as these observations are well placed in time.

## 3.3 Sensitivity with respect to Lag Times $\tau(x)$

For the given application, the calculated lag times $\tau(x)$ refer to the length of time between two medical tests. That is, in a practical application, these lag times give a reference for the time interval between appointments for a patient and his medical practitioner. However, such appointments underlie additional restrictions (such as the disposability of the medical practitioner or a limited flexibility of the patient) which may preclude an exact adherence of the recommended lag times. Especially in South Africa, a limited infrastructure might enforce this situation, which leads back to the sensitivity analysis that was given in section 2.2.4: How do deviations from the optimal lag times increase the expected process costs? This question will now be answered for the resource-constrained setting.

We have shown in Lemma 2.10 that the cost functional $J^u_\lambda$ is continuous with respect to the lag time parameters $\tau(x)$ for all $x \in \mathcal{S}$. In our application, the optimal policy for the resource constrained setting (South Africa) gives only two states ($\big[m,0,0,0\big]$ and $\big[h,0,0,0\big]$) for which diagnostic testing is indicated. We therefore computed the impact of $\tau$-variations around the optimum in Figure 3.3 for the indicated states by using equation (2.16).

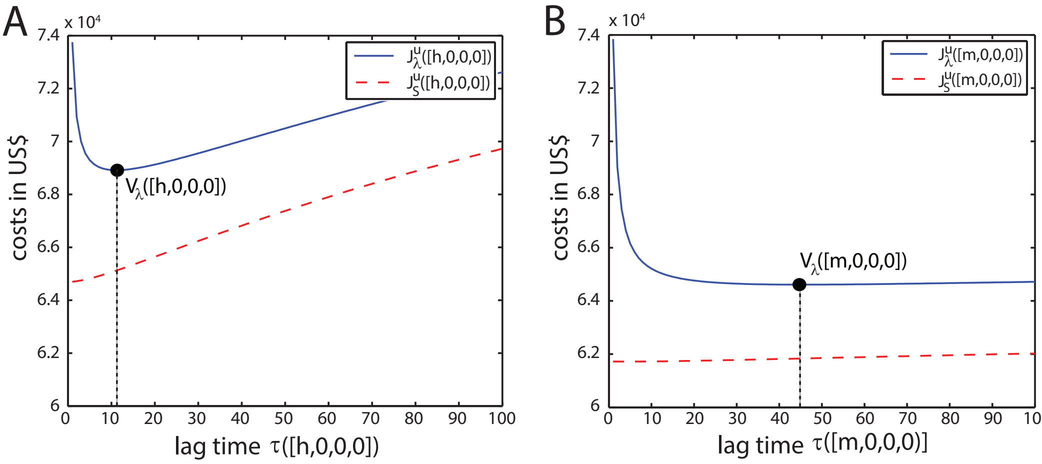

Figure 3.3: **Sensitivity with respect to $\tau$.** Cost functionals $J^u_\lambda(x)$ and $J^u_{\mathcal{S}}(x,u)$ for $x = [h,0,0,0]$ (panel A) and $x = [m,0,0,0]$ (panel B) with $u$ varying in $\tau([h,0,0,0])$ resp. $\tau([m,0,0,0])$ (while being optimal in all other parameters). Taken from [78].

For state $x = \big[h,0,0,0\big]$, the total costs sharply rise if $\tau$ is decreased or -increased (solid blue line in Figure 3.3 A) in relation to its optimum value $\tau^*$ (solid dot). The increase of $J^u_\lambda(x)$ upon increasing values of $\tau$ is paralleled by an increase in the state costs $J^u_{\mathcal{S}}(x)$ (dashed red line). When $\tau$ is decreased, the opposite is true, namely $J^u_{\mathcal{S}}(x)$ decreases, but the overall costs $J^u_\lambda(x)$ increase. Note that the slope of $J^u_{\mathcal{S}}(x)$ (dashed red line) corresponds to the cost-increase attributed to patient health damage.

Although we observe a very sensitive response towards changes in $\tau\left(\left[h,0,0,0\right]\right)$, we get much less sensitivity towards deviations from $\tau^*$ for the second considered state $\left[m,0,0,0\right]$ (see Figure 3.3 B, solid blue line and solid black dot). In particular, upon increases in $\tau$, total- (solid blue line) and state costs (dashed red line) are only marginally increased.

If we focus on the potential health damage to the patient (dashed red lines in Figure 3.3), we can conclude that there is little margin if diagnostic testing is behind schedule for patients that are in state $\left[h,0,0,0\right]$ (high virus load). For state $\left[m,0,0,0\right]$ belated diagnosis will have no rigorous consequences for the health of the patient.

A summary in terms of a two-dimensional contour plot is shown in Figure 3.4. It takes variation in both $\tau\Big(\left[h,0,0,0\right]\Big)$ and $\tau\Big(\left[m,0,0,0\right]\Big)$ simultaneously into account and confirms the observations made from Figure 3.3, indicating that if patients have a high viral load (state $\left[h,0,0,0\right]$), they should strictly comply with the optimal policy.

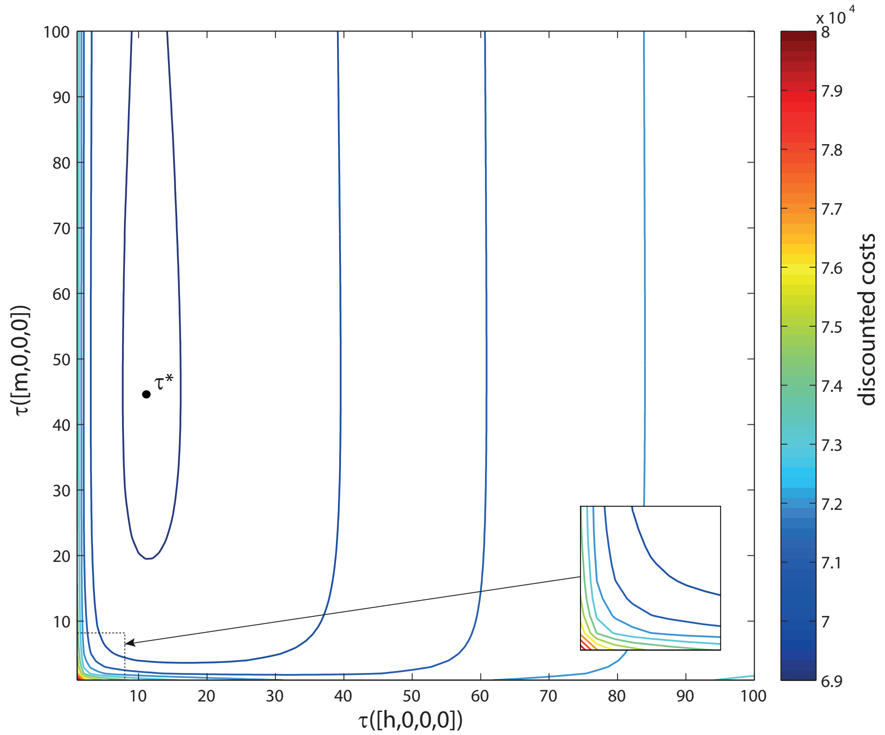

Figure 3.4: **Sensitivity with respect to $\tau$.** Cost functional $J_\lambda(x,u)$ for $x = [h,0,0,0]$ and $u$ varying in $\tau([h,0,0,0])$ (x-axis) and $\tau([m,0,0,0])$ (y-axis), while being optimal in all other parameters. Taken from [78].

## 3.4 Life Expectancy subject to Parameter Variations

We conclude this application chapter by an additional analysis which has no direct analogue in the theoretical part. It is motivated by the special role that the state of death plays in the given application. Besides the minimization of the total expected costs, it is of fundamental interest to maximize the life expectancy of an infected patient. In Table 3.5 and Table 3.6 we already quoted the probability of death after 3, 5, and 15 years for the different settings of control referring to $k_{\text{info}} = \infty$, $k_{\text{info}} = 500$ (resp. $= 400$) and $k_{\text{info}} = 0$. Now, we extend the analysis and see how the probability of death depends on a graduated reduction of the costs for diagnostic tests $k_{\text{info}}$ and the costs of treatment. Instead of the total probability of death, we calculate the probability of AIDS-related death, i.e. we exclude situations of natural death in order to get an unbiased overview of the relations.

The probability of AIDS-related death was computed using a well-known Monte-Carlo method [22]. With baseline parameters, the probability of AIDS-related death for the resource-rich setting after 1000-, 3000- and 5000 days of treatment is 5.0%, 15.5% and 25.2%, see also Figure 3.5 A (blue circles). In the resource-poor setting, the risk of AIDS-related death is higher; 5.1%, 16.3% and 31.0% respectively (see Figure 3.5 C, blue circles), which may be a result of the inability to change treatment in time ($\tau =\geq \tau_{\max}$ for many states in Table 3.3). We therefore evaluated whether a reduction of the test costs may improve survival. In Figure 3.5 A and C we show the probability of AIDS-related death for reduced prices of drug resistance tests $k_{\text{info}} = 200, 100, 50, 5$ US$.
It can be seen that a reduction in diagnostic test prices may significantly improve patient survival in resource-poor settings and that the difference becomes more evident if later time points are evaluated (panel C). For resource-rich countries, patient survival is only insignificantly altered (panel A). To visualize the benefit of reduced diagnostic test prices, we show the 5000 days probability of AIDS-related death as a function of the price reduction factor for drug resistance tests in Figure 3.5 B and D. It can be seen that a price reduction factor of 2.5 (200 US$ per test) in the resource-poor setting may already enable a level of death prevention similar to the resource-rich setting. In the resource-poor setting (panel D) the probability of AIDS-related death 13.7 years (5000 days) after treatment initiation is 31%, 24.2%, 21.8%, 19.2% and 17.6% respectively for tests costs $k_{\text{info}} = 500, 200, 100, 50, 5$ US$ per test. The probability of AIDS-related death 13.7 years (5000 days) after treatment initiation in the resource-rich setting is 25.2%, 24.1%, 22.7%, 21.4% and 20.1% respectively for tests costs $k_{\text{info}} = 400, 200, 100, 50, 5$ US$ per test.

For the resource-rich setting, we also evaluated whether treatment cost reduction would improve patient survival. We found the effect to be quite small: The probability of AIDS-related death 13.7 years (5000 days) after treatment initiation is 23.8%, 23.5%, 23.6% and 21.3% respectively if treatment cost are reduced 2-, 4-, 10-, and 20-fold. It should be mentioned that in resource-rich countries like Germany a more sophisticated treatment schedule may be realistic because more than

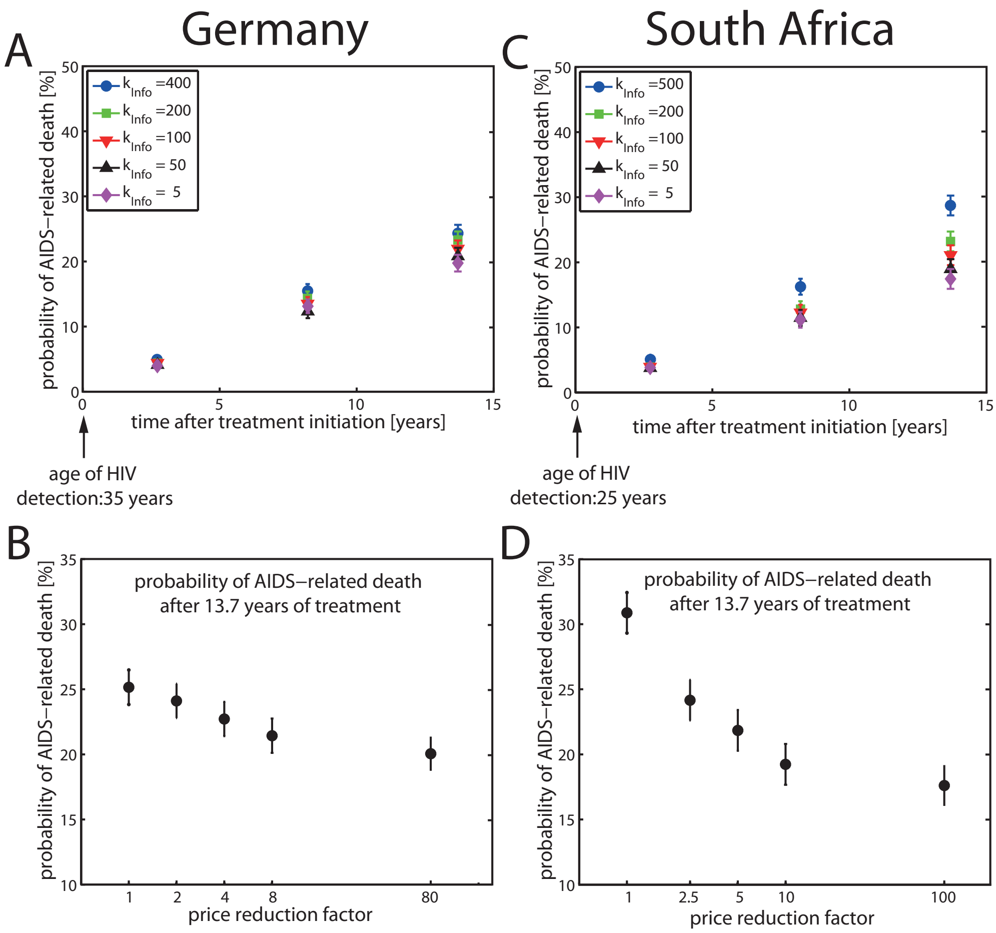

Figure 3.5: **Reduction of diagnostic test costs $k_{\text{info}}$ and effect on AIDS survival** A: Probability of AIDS-related death 1000-, 3000- and 5000 days after treatment initiation under application of a cost-optimal policy with drug resistance test costs $k_{\text{info}} = 400, 200, 100, 50, 5$ US\$ per test in a resource-rich setting (Germany). B: Probability of AIDS-related death 5000 days after treatment initiation as a function of test price reduction in a resource-rich setting (Germany). C: Probability of AIDS-related death 1000-, 3000- and 5000 days after treatment initiation under application of an cost-optimal policy with drug resistance test costs $k_{\text{info}} = 500, 200, 100, 50, 5$ US\$ per test in a resource-constrained setting (South Africa). D: Probability of AIDS-related death 5000 days after treatment initiation as a function of test price reduction in a resource-constrained setting (South Africa). The initial state is given by $x = [h, 0, 0, 0]$.

two treatment options may be available. For resource-poor settings, however, our limited approach is quite realistic.

In conclusion, it may be said that prices for diagnostic test costs are too high in resource-poor settings to allow for cost-optimal **and** life-saving treatment. A small price reduction on the other hand may significantly improve patient survival in a cost-optimal way. In the resource-rich setting it is not indicated that price-reduction for either diagnostics or treatment significantly improve patient survival in a cost-optimal way. The simultaneous price reduction of treatment and diagnostics may, however, do so.

# Summary

For several decades, the theory of Markov decision processes has been successfully used to model situations of controlled stochastic dynamics in various application areas. Beside the original setting which assumes the controlled process to be completely observable at all times, there exist several variants of Markov control theory for cases of incomplete state information. All these variants underlie special restrictions: The time scale of the process has to be discrete, the state space is assumed to exhibit a special ordered structure or the type of dynamics is in some sense predefined. In this thesis we developed a novel model of Markov control with incomplete state information which is applicable to all kinds of continuous-time dynamics on a discrete state space. The observation of the process and the choice of actions take place at discrete points in time which themselves are subject to the control of the decision maker. Each observation produces a fixed amount of information costs $k_{\text{info}}$ which are included in the considered cost criteria. The chosen action determines the stochastic dynamics of the process within the next time period of hidden progress. The resulting combined optimization of observation times and interaction was extensively studied in this thesis.

Given the new setting, we redefined the two criteria of discounted costs and average costs in an appropriate way. Both criteria were analyzed subsequently. Their relation to the cost criteria considered in the original Markov control theory was established and the impact of the control parameters and of the information cost parameter $k_{\text{info}}$ was explored. One main result was the reformulation of the Bellman equation which delivered the basis for an efficient numerical calculation of the optimal control policy. The corresponding value function of optimal costs was discovered to be monotone and continuous in $k_{\text{info}}$ and to coincide with the value function of the original setting when considering vanishing information costs.

The proposed model not only permitted a coherent and productive theoretical analysis, but also formed the basis for an interesting real-world application. We considered the dynamics of HIV and used the developed theory to calculate optimal therapeutic strategies for resource-rich and resource-poor settings. We discovered, among other things, that a decrease of diagnostic costs in resource-poor settings would significantly enhance the medical success of cost optimal therapies.

This thesis provides a comprehensible framework for analyzing situations of controlled dynamics which are not permanently observable. The framework is based on the two fundamental assumptions that a state test always delivers instantaneous and perfect information and that the action can only be adapted after such a test. A question of interest is how far these assumptions can be eased in order to generalize the control model. Finding an answer to this question proves to be a topic of future research.

# Zusammenfassung

Seit einigen Jahrzehnten wird die Theorie der Markov-Kontroll-Prozesse erfolgreich genutzt, um Situationen von kontrollierter stochastischer Dynamik für vielen Anwendungsbereiche zu modellieren. Neben der ursprünglichen Theorie, die von einer vollständigen Beobachtbarkeit des Prozesses ausgeht, existieren verschiedene Ansätze für die Modellierung von unvollständiger Zustandsinformation. Die Ansätze unterliegen alle gewissen Einschränkungen: Der Zeitindex wird als diskret vorausgesetzt, der betrachtete Zustandsraum muss eine spezielle Struktur aufweisen oder die Art der Dynamik ist in einem bestimmten Sinne vordefiniert. In dieser Arbeit wurde eine neuartiges Markov-Kontroll-Modell für den Fall unvollständiger Zustandsinformation entwickelt, das für jegliche kontinuierliche Dynamik auf diskretem Zustandsraum anwendbar ist. Die Beobachtung des Prozesses sowie die Wahl der Kontrollaktionen finden an einzelnen diskreten Zeitpunkten statt, die selbst wiederum der Kontrolle des Entscheiders unterliegen. Jede Beobachtung verursacht fixe Kosten $k_{\text{info}}$, und diese Informationskosten fließen in die betrachteten Kostenkriterien ein. Die gewählte Aktion bestimmt die Entwicklung des Prozesses bis zum Zeitpunkt der nächsten Beobachtung. Das Problem der kombinierten Optimierung von Beobachtungszeitpunkten und Interaktion wurde in dieser Arbeit umfassend erforscht.

Auf Grundlage des neuen Modells wurden die üblichen Kostenkriterien (diskontierte Kosten und langfristige Durchschnittskosten) in geeignetem Sinne formuliert und umfassend analysiert. Dabei wurden ihre strukturellen Eigenschaften betrachtet, der Zusammenhang zu den Kostenkriterien aus der ursprünglichen Theorie wurde erklärt und die Bedeutung der Kontrollparameter und der Informationskosten wurde untersucht. Ein Hauptergebnis war die Umformulierung der Bellman-Gleichung als Basis für eine effiziente numerische Berechnung der optimalen Kontrollstrategie. Außerdem wurde gezeigt, dass die zugehörige Wertefunktion optimaler Kosten monoton und stetig in $k_{\text{info}}$ ist und für verschwindende Informationskosten mit der Wertefunktion der ursprünglichen Theorie übereinstimmt. Das Modell ermöglichte nicht nur eine stimmige und ergebnisreiche theoretische Analyse, sondern war außerdem Ausgangspunkt für ein interessantes Anwendungsbeispiel. Betrachtet wurde die Dynamik des HI-Virus, und die neue Theorie wurde genutzt, um optimale therapeutische Strategien für verschiedene Wirtschaftssituationen zu berechnen. Dabei stellte sich unter anderem heraus, dass eine Senkung der Diagnosekosten in ressourcen-armen Ländern zu einer deutlichen Erhöhung des medizinischen Erfolgs von kostenoptimalen Therapien führen würde.

Diese Arbeit bietet ein anschauliches Modell für die Modellierung von Markov-Kontroll-Prozessen mit beschränkter Zustandsinformation. Das Modell basiert auf den zwei Annahmen, dass eine Beobachtung des Prozesses stets sofortige, perfekte Information liefert und dass die Kontrollaktion nur nach einer solchen Beobachtung angepasst werden kann. Von Interesse ist die Frage, in wieweit diese Annahmen abgeschwächt werden können, um so das Kontrollmodell weiter zu verallgemeinern. Die Beantwortung dieser Frage wird Gegenstand zukünftiger Forschung sein.

# Bibliography

[1] R. F. Anderson and A. Friedman. Optimal inspections in a stochastic control problem with costly observations. *Math. Operations Res.*, 2:155–190, 1977.

[2] A. Arapostatis and V. S. Borkar. A relative value iteration algorithm for non-degenerate controlled diffusions. *SIAM Journal on Control and Optimization*, 50(4):1886–1902, 2012.

[3] M. Avriel. *Nonlinear programming: analysis and methods*. Courier Dover Publications, 2003.

[4] J. Bather. An optimal stopping problem with costly information. *Bulletin of Institute for International Statistics*, 45:9–24, 1973.

[5] R. Bellman. A Markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957.

[6] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[7] D. P. Bertsekas. *Dynamic programming and Optimal Control*. Athena Scientific, 1995.

[8] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[9] A. R. Cassandra, M. L. Littman, and L. P. Kaelbling. Acting optimally in partially observable stochastic domains. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1023–1028, 1994.

[10] H.-T. Cheng. *Algorithms for partially observable Markov decision processes*. PhD thesis, University of British Columbia, Vancouver, 1988.

[11] J. Coffin, F. Maldarelli, S. Palmer, et al. Long-term persistence of low-level HIV-1 in patients on suppressive antiretroviral therapy. Abstract 169, 13th Conference on Retroviruses and Opportunistic Infections; 5–8 February 2006; Denver, Colorado, United States. Available at `http://www.retroconference.org/2006/Abstracts/28061.htm`, accessed May 2012, 2006.

[12] M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM Journal on Numerical Analysis*, 36(2):491–515, 1999.

[13] M. Dellnitz and R. Preis. Congestion and almost invariant sets in dynamical systems. *Proceedings of Symbolic and Numerical Scientific Computation (SNSC'01)*, LNCS 2630:183–209, 2003.

[14] F. D'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10(1):98–108, 1963.

[15] P. Deuflhard. *Newton methods for nonlinear problems*, volume 35 of *Series in Computational Mathematics*. Springer, 2011.

[16] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. Computation of essential molecular dynamics by subdivision techniques I: Basic concept. In *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes in Computational Science and Engineering*, pages 98–115. Springer, 1998.

[17] L. Dieulle, C. Bérenguer, A. Grall, and M. Roussignol. Sequential condition-based maintenance scheduling for a deteriorating system. *European Journal of Operational Research*, 150:451–461, 2003.

[18] N. Djurdjevac, M. Sarich, and C. Schütte. Estimating the eigenvalue error of Markov state models. *Multiscale Modeling and Simulation*, 10(1):61–81, 2012.

[19] J. F. Eastham and K. J. Hastings. Optimal impulse control of portfolios. *Mathematics of Operations Research*, 13(4):588–605, 1988.

[20] J. Fearnley. Exponential lower bounds for policy iteration. *CoRR*, abs/1003.3418, 2010.

[21] D. Finzi, J. Blankson, J. D. Siliciano, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med*, 5(5):512–517, 1999.

[22] D.T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comp Phys*, 22:403–434, 1976.

[23] X. P. Guo and O. Hernández-Lerma. Continuous-time controlled Markov chains. *The Annals of Applied Probability*, 13:363–388, 2003.

[24] X. P. Guo and O. Hernández-Lerma. Continuous-time controlled Markov chains with discounted rewards. *Acta Applicandae Mathematica*, 79:195–216, 2003.

[25] X. P. Guo and O. Hernández-Lerma. Drift and monotonicity conditions for continuous-time controlled Markov chains with an average criterion. *IEEE Transactions on Automatic Control*, 48:236–245, 2003.

[26] X. P. Guo and O. Hernández-Lerma. Continuous-time Markov decision processes: theory and applications. In *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2009.

[27] X. P. Guo and W. P. Zhu. Denumerable state continuous-time Markov decision processes with unbounded cost and transition rates under the discounted criterion. *Journal of Applied Probability*, 39:233–250, 2002.

[28] E. A. Hansen. An improved policy iteration algorithm for partially observable MDPs. *Advances in Neural Information Processing Systems*, 10, 1998.

[29] E. A. Hernandez-Vargas, R. H. Middleton, and P. Colaneri. Optimal and MPC switching strategies for mitigating viral mutation and escape. In *Preprints of the 18th IFAC World Congress Milano (Italy) August 28- September 2*, 2011.

[30] O. Hernández-Lerma and S. I. Marcus. Adaptive control of Markov processes with incomplete state information and unknown parameters. *Jurnal of Optimization Theory and Applications*, 52, 1987.

[31] R. Hollanders, J.-C. Delvenne, and R. M. Jungers. The complexity of policy iteration is exponential for discounted Markov decision processes. Available at `http://perso.uclouvain.be/romain.hollanders/docs/CDC12_HollandersDelvenneJungers.pdf`, accessed Dec. 2012.

[32] R. A. Howard. *Dynamic programming and Markov processes*. MIT Press, 1960.

[33] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[34] M. J. Kim. *Optimal Control and Estimation of Stochastic Systems with Costly Partial Information*. PhD thesis, University of Toronto, 2012.

[35] R. L. A. Kirch and M. Klein. Surveillance schedules for medical examinations. *Management Science*, 20(10):1403–1409, 1974.

[36] R. Korn. Portfolio optimisation with strictly positive transaction costs and impulse control. *Finance and Stochastics*, 2(2):85–114, 1998.

[37] R. Korn. Some applications of impulse control in mathematical finance. *Mathematical Methods of Operations Research*, 50(3):493–518, 1999.

[38] Y. Kuo. Optimal adaptive control policy for joint machine maintenance and product quality control. *European Journal of Operational Research*, 171, 2006.

[39] E. G. Kyriakidis. Optimal control of a simple immigration birth death process through total catastrophes. *European Journal of Operational Research*, 81, 1995.

[40] O. Lambotte, M.-L. Chaix, B. Gubler, et al. The lymphocyte HIV reservoir in patients on long-term HAART is a memory of virus evolution. *AIDS*, 18(8):1147–1158, 2004.

[41] J. B. Lasserre and O. Hernández-Lerma. *Discrete-time Markov control processes: basic optimality criteria*. Springer, 1996.

[42] C. Lefèvre. Optimal control of a birth and death epidemic process. *Operations Research*, 29, 1981.

[43] M. Littman, T. Dean, and L. Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 394–402. Morgan Kaufmann Publishers Inc., 1995.

[44] R. Luo, M. J. Piovoso, J. Martinez-Picado, and R. Zurakowski. Optimal antiviral switching to minimize resistance risk in HIV therapy. *PLoS One*, 6(11):e27047, 2011.

[45] R. C. Merton. *Optimum consumption and portfolio rules in a continuous-time model*. MIT, 1970.

[46] G. E. Monahan. A survey of partially observable Markov decision processes: theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

[47] E. Pashenkova, I. Rish, and R. Dechter. Value iteration and policy iteration algorithms for Markov decision problem. Available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.8155`, accessed Dec. 2012.

[48] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[49] D. Rhenius. Incomplete information in Markovian decision models. *The Annals of Statistics*, 2:1327–1334, 1974.

[50] D. B. Rosenfield. Markovian deterioration with uncertain information. *Operations Research*, 24:141–155, 1976.

[51] S. M. Ross. Quality control under Markovian deterioration. *Management Science*, 17:587–596, 1971.

[52] I. M. Rouzine, A. Rodrigo, and J. M. Coffin. Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiol Mol Biol Rev*, 65(1):151–185, 2001.

[53] M. Sarich, F. Noé, and C. Schütte. On the approximation quality of Markov state models. *Multiscale Modeling and Simulation*, 8(4):1154–1177, 2010.

[54] I. R. Savage. Surveillance problems. *Naval Research Logistics*, 9:187–209, 1962.

[55] Y. Sawaragi and T. Yoshikawa. Discrete-time Markovian decision processes with incomplete state observation. *Annals of Mathematical Statistics*, 41:78–86, 1970.

[56] C. Schütte and W. Huisinga. Markov processes on discrete state spaces. Available at `http://numerik.mi.fu-berlin.de/wiki/WS2012/Seminare/Seminar_Stochastik_Dokumente/SchuetteHuisinga_MarkovChainsOnDiscreteStateSpace.pdf`, accessed Dec. 2012.

[57] C. Schütte, S. Winkelmann, and C. Hartmann. Optimal control of molecular dynamics using Markov state models. *Math. Program. (Series B)*, 134:259–282, 2012.

[58] P. Sendi, H. F. Günthard, M. Simcock, et al. Cost-effectiveness of genotypic antiretroviral resistance testing in HIV-infected patients with treatment failure. *PLoS One*, 2(1):e173, 2007.

[59] L. I. Sennott. The convergence of value iteration in average cost Markov decision chains. *Operations Research Letters*, 19, 1996.

[60] R. Serfozo. Optimal control of random walks, birth and death processes and queues. *Advances in Applied Probability*, 13, 1981.

[61] S. M. Shechter, M. D. Bailey, and A. J. Schaefer. A modeling framework for replacing medical therapies. *IIE Transactions*, 40:861–869, 2008.

[62] E. J. Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford Univerity, 1971.

[63] E. J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Operations Research*, 26:282–304, 1978.

[64] M. Stoll, C. Kollan, F. Bergmann, et al. Calculation of direct antiretroviral treatment costs and potential cost savings by using generics in the German HIV ClinSurv cohort. *PLoS One*, 6(9):e23946, 2011.

[65] The Clinton Health Access Initiative. Antiretroviral (ARV) ceiling price list. Available at `http://www.clintonfoundation.org`, accessed May 2012.

[66] The International Monetary Fund. World economic outlook database. Available at `http://www.imf.org`, accessed May 2012.

[67] UNAIDS. Report on the global AIDS epidemic 2010. Available at `http://www.unaids.org/documents/20101123_globalreport_em.pdf`, accessed May 2012.

[68] M. von Kleist, S. Menz, and W. Huisinga. Drug-class specific impact of antivirals on the reproductive capacity of HIV. *PLoS Comput Biol*, 6:e1000720, 2010.

[69] M. von Kleist, S. Menz, H. Stocker, K. Arasteh, Ch. Schütte, and W. Huisinga. HIV quasispecies dynamics during pro-active treatment switching: Impact on multi-drug resistance and resistance archiving in latent reservoirs. *PLoS One*, 6:e18204, 2011.

[70] T. Wang, P. Poupart, M. Bowling, and D. Schuurmans. Compact, convex upper bound iteration for approximate POMDP planning. *International Journal of Production Research*, 38:1425–1436, 2000.

[71] D. Werner. *Funktionalanalysis.* Springer, 2005.

[72] C. White. Optimal inspection and repair of a production process subject to deterioration. *Journal of the Operational Research Society*, 29:235–243, 1978.

[73] D. J. White. Dynamic programming, Markov chains, and the method of successive approximations. *Math. Anal. Appl.*, 6:373–376, 1963.

[74] D. J. White. A survey of applications of Markov decision processes. *J Opl Res Soc*, 44:1073–1096, 1993.

[75] D. J. White. *Markov Decision Processes.* John Wiley and Sons, 1993.

[76] D. V. Widder. *The Laplace Transform.* Princeton University Press, 1946.

[77] S. Winkelmann, C. Schütte, and Max von Kleist. Markov control processes with rare state observation: Theory and application to treatment scheduling in HIV-1. *Comm. Math. Sci.*, submitted, 2012.

[78] S. Winkelmann, C. Schütte, and Max von Kleist. Markov control with rare state observation: Sensitivity analysis with respect to optimal treatment strategies against HIV-1. *International Journal of Biomathematics and Biostatistics*, submitted, 2012.

[79] Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36:593–603, 2011.

[80] M. Zelen. Optimal scheduling of examinations for the early detection of disease. *Biometrica*, 80:279–293, 1993.

[81] Q. Zhu. Average optimality for continuous-time Markov decision processes with a policy iteration approach. *Journal of Mathematical Analysis and Applications*, 339:691–704, 2008.

[82] Q. Zhu, X. Yang, and Ch. Huang. Policy iteration for continuous-time average reward Markov decision processes in Polish spaces. *Abstract and Applied Analysis*, 2009.

# List of Symbols

| | |
|---|---|
| $\mathcal{A}$ | action space, 6, 44 |
| $\mathcal{A}(x) \subset \mathcal{A}$ | set of available actions for state $x \in \mathcal{S}$, 6, 44 |
| $A_t$ | action applied at time $t$, 45 |
| $a : \mathcal{S} \to \mathcal{A}$ | part of the policy $u = (a, \tau)$, 45 |
| $a \in \mathcal{A}$ | action, 7 |
| $\mathcal{B}(\mathcal{A})$ | Borel $\sigma$-algebra on $\mathcal{A}$, 6 |
| $C : \mathcal{S} \times \mathcal{A} \times (0, \infty] \to [0, \infty)$ | function of expected discounted costs depending on the initial state $x \in \mathcal{S}$, the action $a \in \mathcal{A}$ and the lag time $\tau \in (0, \infty]$, 49 |
| $C_u(x) = C(x, a(x), \tau(x))$ | function of expected discounted costs for a given policy $u \in \mathcal{U}$, 51 |
| $\tilde{c} : \mathcal{S} \times \mathcal{A} \to [0, \infty)$ | cost function for a discrete-time Markov decision process, 11 |
| $c : \mathcal{S} \times \mathcal{A} \to [0, \infty)$ | cost function, 7, 44 |
| $c_{\mathcal{A}} : \mathcal{A} \to [0, \infty)$ | part of the cost function $c(x, a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$, 9 |
| $c_{\mathcal{S}} : \mathcal{S} \to [0, \infty)$ | part of the cost function $c(x, a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$, 9 |
| $D_\tau$ | diagonal matrix with $D_\tau(x, x) = e_\tau(x)$, 51 |
| $\delta_x$ | Dirac measure on $\mathcal{S}$, 8 |
| $\mathbb{E}_\nu^\pi$, $\mathbb{E}_\nu^u$ | expectation value w.r.t. $\mathbb{P}_\nu^\pi$ resp. $\mathbb{P}_\nu^u$, 8, 45 |
| $\mathbb{E}_x^\pi$, $\mathbb{E}_x^u$ | expectation value w.r.t. $\mathbb{P}_x^\pi$ resp. $\mathbb{P}_x^u$, 8, 45 |
| $E$ | matrix choosing the first component of a vector, 21, 87 |
| $e_\tau$ | vector with entries $e_\tau(x) = e^{-\lambda\tau(x)}$, 51 |
| $\eta_u$ | constant of expected average costs for a given policy $u \in \mathcal{U}$, 20 |
| $\tilde{\eta}$ | constant of expected average costs for a discrete-time Markov control process, 11 |
| $I$ | identity matrix, 14 |
| $J_\lambda : \mathcal{S} \times \mathcal{U} \to [0, \infty)$ | function of expected discounted costs for a discount factor $\lambda > 0$, 10, 46 |
| $\tilde{J}_\lambda : \mathcal{S} \times \mathcal{U} \to [0, \infty)$ | function of expected discounted costs for a discrete-time Markov control process, 11 |
| $J_\lambda^u$ | function of expected discounted costs for a fixed policy $u$, 50 |
| $\bar{J} : \mathcal{S} \times \mathcal{U} \to [0, \infty)$ | function of expected average costs, 10, 47 |

| | |
|---|---|
| $\tilde{J} : \mathcal{S} \times \mathcal{U} \to [0, \infty)$ | function of expected average costs for a discrete-time Markov control process, 11 |
| $k_{\text{info}}$ | information costs, 44 |
| $L_a$ | generator for action $a \in \mathcal{A}$, 6, 44 |
| $L_u(x, y) = L_{u(x)}(x, y)$ | adapted generator for a given policy $u \in \mathcal{U}$, 14 |
| $l_a(x) = -L_a(x, x)$ | jump time parameter for a given action $a \in \mathcal{A}$, 6, 44 |
| $l_u(x) = -L_{u(x)}(x, x)$ | jump time parameter for a given policy $u \in \mathcal{U}$, 12 |
| $\lambda > 0$ | discount factor, 10 |
| $\mu_u$ | equilibrium distribution for a given policy $u \in \mathcal{U}$, 20 |
| $\nu$ | initial distribution, 8 |
| $\mathbb{P}^\pi_\nu$, $\mathbb{P}^u_\nu$ | probability measure on the set of possible state-action-realizations for an initial distribution $\nu$ and a policy $\pi$ resp. $u$, 8, 45 |
| $\mathbb{P}^\pi_x$, $\mathbb{P}^u_x$ | probability measure on the set of possible state-action-realizations for an initial state $x$ and a policy $\pi$ resp. $u$, 8, 45 |
| $P_a$ | transition matrix for a discrete-time Markov control process, 11 |
| $P_{a,\tau}$ | transition matrix given action $a \in \mathcal{A}$ and lag time $\tau \in (0, \infty]$, 50 |
| $P_u$ | transition matrix for a policy $u \in \mathcal{U}$, 50 |
| $\Pi$ | set of randomized, time-dependent Markov policies, 7 |
| $(\pi_t)_{t \geq 0}$ | randomized, time-dependent Markov policy, 7 |
| $\mathcal{S}$ | state space, 6, 44 |
| $T^* : \mathbb{R}^{\lvert\mathcal{S}\rvert} \to \mathbb{R}^{\lvert\mathcal{S}\rvert}$ | operator, 18 |
| $T_a : \mathbb{R}^{\lvert\mathcal{S}\rvert} \to \mathbb{R}^{\lvert\mathcal{S}\rvert}$ | operator for a given action $a \in \mathcal{A}$, 13 |
| $T_u : \mathbb{R}^{\lvert\mathcal{S}\rvert} \to \mathbb{R}^{\lvert\mathcal{S}\rvert}$ | operator for a given policy $u \in \mathcal{U}$, 13, 50 |
| $T_{a,\tau}$ | operator for a given action $a \in \mathcal{A}$ and a lag time $\tau \in (0, \infty]$, 50 |
| $(t_j)_{j \in \mathbb{N}_0}$ | observation times, 44 |
| $t \geq 0$ | continuous time index, 7 |
| $\tau : \mathcal{S} \to (0, \infty]$ | part of the policy $u = (a, \tau)$, 45 |
| $\tau > 0$ | lag time, 44 |
| $\mathcal{U}$ | set of deterministic stationary Markov policies, 7 |
| $u : \mathcal{S} \to \mathcal{A} \times (0, \infty]$ | deterministic stationary Markov policy in the setting of information costs, 45 |
| $u : \mathcal{S} \to \mathcal{A}$ | deterministic stationary Markov policy in the original setting, 7 |
| $u^* \in \mathcal{U}$ | optimal policy, 10 |
| $\bar{V} : \mathcal{S} \to [0, \infty)$ | value function of average costs, 10, 47 |
| $V_\lambda : \mathcal{S} \to [0, \infty)$ | value function of discounted costs for a discount factor $\lambda > 0$, 10, 46 |
| $X_t$ | state of the process at time $t$, 45 |