Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

STEFANIE WINKELMANN[1]

Markov Control with Rare State Observation: Average Optimality [2]

[1]Zuse Institute Berlin (ZIB), Germany
[2]to appear in: 2017

ZIB Report 16-59 (November 2016)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Markov Control with Rare State Observation: Average Optimality

Stefanie Winkelmann[‡]

Abstract

This paper investigates the criterion of long-term average costs for a Markov decision process (MDP) which is not permanently observable. Each observation of the process produces a fixed amount of *information costs* which enter the considered performance criterion and preclude from arbitrarily frequent state testing. Choosing the *rare* observation times is part of the control procedure. In contrast to the theory of partially observable Markov decision processes, we consider an arbitrary continuous-time Markov process on a finite state space without further restrictions on the dynamics or the type of interaction. Based on the original Markov control theory, we redefine the control model and the average cost criterion for the setting of information costs. We analyze the constant of average costs for the case of ergodic dynamics and present an optimality equation which characterizes the optimal choice of control actions and observation times. For this purpose, we construct an equivalent freely observable MDP and translate the well-known results from the original theory to the new setting.

Key words: Markov decision processes, partial observability, information costs, average optimality.
MSC2000 subject classification: 90C40, 90C39, 49N30

1 Introduction

Markov decision processes (MDP's) provide a mathematical framework for modeling situations in which the dynamics of a process are partly random and partly under the control of a decision maker. Such situations appear in many application areas like machine maintenance, portfolio management or medical therapy. Given the state of the process at some point in time, the decision maker has to choose an action which affects the future dynamics of the process. Depending on the process evolution the system produces costs (or rewards), and the goal is to optimize the dynamics according to a given performance criterion. Bellman [4, 5] and Howard [19] popularized the theory of sequential decision making in the 1960's. They introduced the dynamic programming concept which is expressed in the so called *Bellman equation* and states the optimization problem in a recursive form.

Since then, MDP's are an object of steady research, see e.g. [9–12, 14, 17, 18, 26, 31, 32] where different variants with respect to the considered time index, the state space or the performance criterion are studied. Of special interest is the situation of limited state information. How should the action be chosen if the state of the process is not known with certainty? These

[‡]Zuse Institute Berlin (ZIB), Germany, winkelmann@zib.de

situations are considered in the theory of partially observable MDP's, see the survey paper by Monahan [24] for an overview. Most of the approaches assume the underlying time index to be discrete, see [20, 22, 29]. A continuous-time partially observable process is considered in [1–3,21,28]; however, these approaches are based on restrictions with respect to the considered dynamics and/or the kind of interaction: The system is assumed to be non-decreasing in the state space (which refers to considering a process of deterioration), interaction consists of stopping the process or bringing it back to some initial state.

In this paper we avoid such restrictions and consider an arbitrary continuous-time Markov decision process on a finite state space which is not permanently observable. Instead, each observation of the process produces a fixed amount of *information costs*

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

which are included in the considered performance criterion and preclude from arbitrarily frequent state testing. To determine the dates for the *rare* but *flexible* observations is part of the control procedure. Given a state observation, the decision maker chooses both a time for the next observation as well as an action which determines the stochastic dynamics within the subsequent time interval of hidden progress. Within such a time interval the action is fixed, i.e. it cannot be changed blindly without knowing the state.

The approach is motivated by the fact that in many real-world applications a permanent observation and control of the process under consideration is not feasible. Especially situations in the context of medical therapy or machine maintenance suggest that state examinations are costly and therefore rare. Choosing suitable dates for any kind of inspections is a common task to perform: A medical scientist proposes a date for the next checkup, and a machinist fixes a date for the next inspection of the machine. Given the result of an inspection, a decision concerning future interaction (medical treatment of a patient, choice of production mode for the machine...) has to be taken. The model presented here exactly reflects this situation of rare but flexible observation and decision making.

The setting of Markov control with information costs has been investigated in [33,34] and [8] for the criterion of discounted costs over an infinite time horizon. In the present paper, we will analyze the criterion of *long-term average costs* which - in the context of completely observable MDP's - is a common optimality criterion studied by many authors, see e.g. [7,16,23,26,27,35]. We will redefine the performance criterion for the given situation of information costs (Section 2) and derive the corresponding Bellman equation. Interestingly, the approach will clearly differ from the one given in [8,33,34]: 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 MDP which will be done in Section 3. A further cost analysis concerning the impact of the information cost parameter k_{info} is given in Section 4.

2 The control model

The control model considered in this article is given by the tuple

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

where $\mathcal{S}$ and $\mathcal{A}$ are the *state space* and the *action space*, respectively. While the state space is assumed to be finite, the action space may be any metrizable topological space; we denote

the corresponding Borel σ-algebra by $\mathcal{B}(\mathcal{A})$. It is $\mathcal{A}(x)$ the set of *available actions* when the process is in state $x \in \mathcal{S}$. 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 $\sup_{a \in \mathcal{A}(x)} |L_a(x,x)| < \infty \quad \forall x \in \mathcal{S}$. It is $c : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ the *cost function* giving the costs produced by the process per unit of time depending on the actual state and the chosen action. Furthermore, - in contrast to the original Markov control theory - this model contains the parameter k_{info} of **information costs** that have to be paid each time that the process is observed.

2.1 The controlled process

The corresponding Markov control process $(X_t)_{t \geq 0}$ - which is itself continuous in time - may only be observed at *discrete but flexible* points in time which themselves are subject to the control of the decision maker. More precisely, the **control procedure** is the following. Starting with some (known) state $X_{t_0} = x_0 \in \mathcal{S}$ at time $t_0 = 0$, the decision maker 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, one can only determine the state X_{t_1} at time t_1 by making a test which produces costs k_{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.*

In this setting, the state tests are assumed to give instantaneous and full information such that the state of the process at the observation times is known with certainty. Moreover, we assume that 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. A control policy in this setting is given by a function

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

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

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

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

assuming that this limit exists; otherwise we set $e^{L_a \infty} := I$. (In fact, the definition of $e^{L_a \infty}$ is of no further significance as this expression will always be multiplied by $e^{-\lambda\infty} = 0$.) 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.

The observation times $(t_j)_{j\in\mathbb{N}_0}$ given by $t_0=0$ and $t_{j+1}=t_j+\tau(X_{t_j})$ are random variables that depend on the process evolution and may take on the value ∞; in this regard we interpret $t_j+\infty:=\infty$ for $t_j<\infty$ as well as $\infty+\infty:=\infty$.

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

$$\left\{(X_t,A_t)_{t\geq 0}\colon X_t\in\mathcal{S}, A_t\in\mathcal{A}\ \forall t\geq 0\right\}$$

with observation times $(t_j)_{j\in\mathbb{N}_0}$ by

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

We write $\mathbb{P}^u_x$ for $\nu=\delta_x$ (deterministic start in $x\in\mathcal{S}$) and denote the corresponding expectation values by $\mathbb{E}^u_\nu$ resp. $\mathbb{E}^u_x$.

Figure 1 illustrates the controlled process $(X_t)_{t\geq 0}$ for a fixed policy $u\in\mathcal{U}_{\text{info}}$.

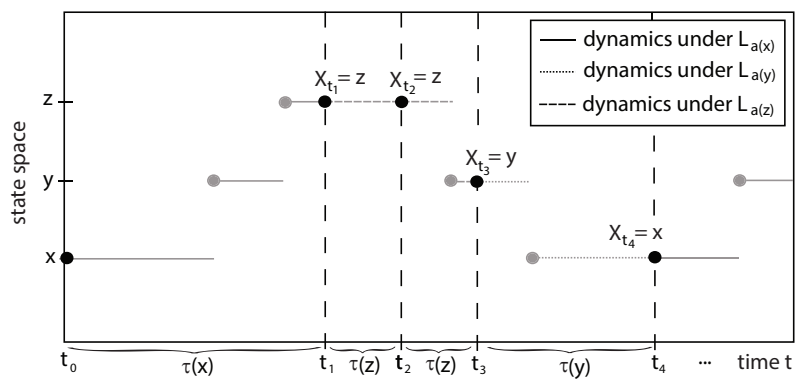

Figure 1: **Controlled Markov process with information costs.** Possible trajectory of a rarely observable MDP given a deterministic stationary policy u. Starting in a (known) state $x\in\mathcal{S}$ at time t_0 the dynamics of the process during the time interval $[t_0,t_0+\tau(x))$ are determined by the generator $L_{a(x)}$, i.e. the process stays in x for some random period of time which is exponentially distributed with parameter $-L_{a(x)}(x,x)$ and then jumps to a state $y\neq x$ with probability $\frac{L_{a(x)}(x,y)}{-L_{a(x)}(x,x)}$. 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.

2.2 Criterion of long-run average costs

Within the original Markov control theory, the criterion of long-run average costs is one of the most studied performance criteria. It is not self-evident how the existing definitions of this cost criterion can be transferred to the novel setting of information costs because the observation times and information costs have to be weighted in an suitable way. An adequate reformulation is the following.

Definition 2.1 (The expected average cost criterion)**.** Given an initial state $x \in \mathcal{S}$, the long-run expected average costs under control $u \in \mathcal{U}_{\text{info}}$ are defined by

$$J(x,u) := \limsup_{T\to\infty} \mathbb{E}_x^u \left(\frac{1}{T} \sum_{\substack{j\in\mathbb{N}_0 \\ t_j<T}} \left(\int_{t_j}^{t_{j+1}\wedge T} c(X_s, a(X_{t_j}))\, ds + k_{\text{info}} \right) \right), \tag{2.2}$$

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

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

Note that in (2.2) 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 cannot be observed during such an interval) changes as usual, while the action stays the same.

Remark 2.2. *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* (2.2) *as*

$$J(x,u) = \limsup_{n\to\infty} \mathbb{E}_x^u \left(\frac{1}{t_n} \sum_{j=0}^{n-1} \left(\int_{t_j}^{t_{j+1}} c(X_s, a(X_{t_j}))\, ds + k_{\text{info}} \right) \right). \tag{2.3}$$

For infinite t_j, however, this expression has no direct interpretation, which motivates to choose the more general notation given in equation (2.2)*.*

Under certain conditions - which will be proposed now - the function $J(x,u)$ of average costs does not depend on the state x but is given by a constant.

Ergodic dynamics and finite lag times.

In the setting of information costs, the dynamics of the controlled process are named to be *ergodic* if the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ is irreducible. Note that this property only depends on the part of the policy $u(x) = (a(x), \tau(x))$ declaring the action $a(x)$ for each state $x \in \mathcal{S}$, while the lag times $\tau(x)$ are irrelevant.

For the case of ergodic dynamics, we tend to express the cost functional $J(x,u)$ for a given policy $u \in \mathcal{U}_{\text{info}}$ with **finite** lag times (i.e. $\tau(x) < \infty$ for all $x \in \mathcal{S}$) in terms of a stationary distribution of the process $(X_t)_{t\geq 0}$. However, the controlled process $(X_t)_{t\geq 0}$ itself is <u>not</u> 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 a stationary distribution could be characterized.
Being aware that also the costs produced by the process at time t depend on the 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

$$P_u(x,y) = e^{L_{a(x)}\tau(x)}(x,y) \quad \text{for all } x,y \in \mathcal{S}. \tag{2.4}$$

Let us denote the stationary distribution with respect to P_u by μ, i.e. we assume $\mu \in \mathbb{R}^{|\mathcal{S}|}$ to be a probability vector with $\mu P_u = \mu$. For ergodic dynamics this stationary distribution is unique. 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

$$C(x,a(x),\tau(x)) := \mathbb{E}_x^{a(x)}\left(\frac{1}{\tau(x)}\int_0^{\tau(x)} c(X_s,a(x))\,ds\right). \tag{2.5}$$

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

$$\tilde{C}(x,a(x),\tau(x)) := C(x,a(x),\tau(x)) + \frac{k_{\text{info}}}{\tau(x)}. \tag{2.6}$$

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 stationary 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. we set

$$\tilde{\mu}(x) = \frac{\mu(x)\tau(x)}{\sum_{y\in\mathcal{S}} \mu(y)\tau(y)}. \tag{2.7}$$

Lemma 2.3. *For a policy $u \in \mathcal{U}_{\text{info}}$ with finite lag times and unique stationary distribution μ (fulfilling $\mu P_u = \mu$) the long-term average costs are given by the $\tilde{\mu}$-weighted average of $\tilde{C}$, i.e. it holds*

$$J(x,u) = \sum_{y\in\mathcal{S}} \tilde{\mu}(y)\tilde{C}(y,u(y)) = \langle \tilde{\mu}, \tilde{C}_u \rangle =: \eta_u \quad \textit{for all } x \in \mathcal{S} \tag{2.8}$$

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

Proof. As the state space is assumed to be finite and the policy is homogeneous in time, the cost function $c(X_s, a(X_{t_j}))$ appearing in (2.3) is bounded, such that, by the dominated convergence theorem, we can take the limit into the expectation and get

$$J(x,u) = \mathbb{E}_x^u\left(\limsup_{n\to\infty} \frac{1}{t_n}\sum_{j=0}^{n-1}\left(\int_{t_j}^{t_{j+1}} c(X_s,a(X_{t_j}))\,ds + k_{\text{info}}\right)\right). \tag{2.9}$$

We rewrite $t_n = \sum_{j=0}^{n-1} \tau(X_{t_j})$ and apply the strong law of large numbers to find

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

as well as

$$\sum_{j=0}^{n-1} \left(\int_{t_j}^{t_{j+1}} c(X_s, a(X_{t_j}))\, ds + k_{\text{info}} \right) \sim n \sum_{y \in \mathcal{S}} \mu(y) \cdot \tau(y) \cdot \tilde{C}(y, a(y), \tau(y)) \quad (a.s.) \quad \text{for } n \to \infty.$$

Inserting into (2.9) delivers

$$\begin{aligned} J(x,u) &= \mathbb{E}_x^u \left(\sum_{y \in \mathcal{S}} \frac{\mu(y)\tau(y)}{\sum_{z \in \mathcal{S}} \mu(z)\tau(z)} \cdot \tilde{C}(y, a(y), \tau(y)) \right) \\ &= \mathbb{E}_x^u \left(\sum_{y \in \mathcal{S}} \tilde{\mu}(y) \cdot \tilde{C}(y, a(y), \tau(y)) \right) \\ &= \sum_{y \in \mathcal{S}} \tilde{\mu}(y) \cdot \tilde{C}(y, a(y), \tau(y)) \end{aligned}$$

independent of x, which gives (2.8). □

3 Average optimality

In this section we present our main result, Theorem 3.1, which states the analogy between a given MDP with information costs and a freely observable MDP. By means of this analogy it will be possible to directly reformulate the central Bellman equation of average optimality for the setting of rare state observation.

3.1 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 $u \in \mathcal{U}_{\text{info}}$. The idea is to formulate another Markov 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 τ, this residence time can be any multiple of τ. The number of time intervals of length τ 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

$$G_{a,\tau}(x,y):=\frac{1}{\tau}e^{L_a\tau}(x,y)\quad \text{for } y\neq x,\quad G_{a,\tau}(x,x):=-\frac{1}{\tau}\left(1-e^{L_a\tau}(x,x)\right). \tag{3.10}$$

For $\tau=\infty$ we set

$$G_{a,\infty}(x,y):=0\quad \forall y\in\mathcal{S} \tag{3.11}$$

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

$$\lim_{\tau\to\infty} G_{a,\tau}=G_{a,\infty} \tag{3.12}$$

which will be relevant for future statements.

The information process $(Y_t)_{t\geq 0}$.

Now, we define 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: τ 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

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

In terms of the transition matrix P_u defined in (2.4) it holds, as for finite $\tau(x)$,

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

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 information 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 (2.6) 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

$$\tilde{C}(x,a,\infty)=C(x,a,\infty):=\lim_{T\to\infty}\mathbb{E}_x^a\left(\frac{1}{T}\int_0^T c(X_s,a)\,ds\right). \tag{3.13}$$

Now we can formulate our main result: the analogy of the processes $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ with respect to the average cost criterion.

Theorem 3.1. *For each policy $u\in\mathcal{U}_{\text{info}}$ the freely observable MDP $(Y_t)_{t\geq 0}$ with generators $G_{a,\tau}$ and cost function $\tilde{C}$ has the same expected average costs as the rarely observable MDP $(X_t)_{t\geq 0}$, i.e. it holds*

$$\lim_{T\to\infty}\mathbb{E}_x^u\left(\frac{1}{T}\int_0^T \tilde{C}(Y_s,u(Y_s))\,ds\right)=J(x,u) \tag{3.14}$$

with $J(x,u)$ defined in (2.2).

Note that in (3.14) the second argument $u(Y_s)$ of the cost function $\tilde{C}$ is running in time, whereas in the definition (2.2) of $J(x,u)$ it was evaluated at the beginning t_j of a time interval.

Proof of Theorem 3.1. For ergodic dynamics and finite lag times we only need to show that $\tilde{\mu}$ (defined in (2.7)) is the stationary distribution for the process $(Y_t)_{t\geq 0}$, which simply follows from

$$\tilde{\mu}G_u=\frac{1}{\sum_{y\in\mathcal{S}}\tau(y)\mu(y)}(\mu P_u-\mu I)=0$$

because $\mu P_u=\mu$. Then it holds

$$\lim_{T\to\infty}\mathbb{E}_x^u\left(\frac{1}{T}\int_0^T \tilde{C}(Y_s,u(Y_s))\,ds\right)=\sum_{y\in\mathcal{S}}\tilde{\mu}(y)\tilde{C}(y,u(y))=J(x,u)\quad\forall x\in\mathcal{S}, \tag{3.15}$$

compare (2.8).
In the case of non-ergodic dynamics under the policy $u\in\mathcal{U}_{\text{info}}$, let $\mathcal{S}=C_1\cup\ldots\cup C_d\cup D$ be the split-up of the state space for the observation process $(X_{t_j})_{j\in\mathbb{N}}$ into d mutually disjoint communication classes $C_1,\ldots,C_d$ and a set D of transient states $D=\{x\in\mathcal{S}:\pi(x)=0\}$ where π is the corresponding maximal stationary distribution on $\mathcal{S}$, see [6, 25] for background details. For each class C_k there exists a stationary distribution μ_k on $\mathcal{S}$ with $\mu_k P_u=\mu_k$ such that $C_k=\{x\in\mathcal{S}:\mu_k(x)>0\}$. We first assume all lag times to be finite which implies that the same split-up holds for the underlying process $(X_t)_{t\geq 0}$. Moreover, by defining $\tilde{\mu}_k(x):=\frac{\mu_k(x)\tau(x)}{\sum_{y\in\mathcal{S}}\mu_k(y)\tau(y)}$ we obtain for each $k=1,\ldots,d$ a stationary distribution of the process $(Y_t)_{t\geq 0}$ with $\tilde{\mu}_k(x)>0$ if and only if $\mu_k(x)>0$. Thereby, also the space decompositions of the processes $(X_{t_j})_{j\in\mathbb{N}}$ and $(Y_t)_{t\geq 0}$ agree. For each initial state x belonging to a class C_k the long-term average costs for

both $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ are given by the $\tilde{\mu}_k$-weighted average of the costs function $\tilde{C}$, compare Equation (3.15) with $\tilde{\mu}$ replaced by $\tilde{\mu}_k$. Let η_k denote this constant of average costs for the class C_k. For an initial state $x \in D$, the long-term average costs $J(x,u)$ for the process (X_t) are given by a weighted average of the constants η_k,

$$J(x,u) = \sum_{k=1}^{d} q_k(x)\eta_k$$

with the weight $q_k(x) := \mathbb{P}_x^u(X_t \in C_k \text{ for some } t > 0)$ referring to the probability to end up in class C_k after starting in x. The weights are characterized by the recursion

$$\begin{aligned} q_k(x) &= \sum_{y\in\mathcal{S}} \mathbb{P}_x^u(X_t \in C_k \text{ for some } t > \tau(x) | X_{\tau(x)} = y) \cdot \mathbb{P}_x^u(X_{\tau(x)} = y) \\ &= \sum_{y\in C_k} P_u(x,y) + \sum_{z\in D} P_u(x,z) q_k(z) \end{aligned} \tag{3.16}$$

resulting from the law of total probability with respect to the state $X_{\tau(x)}$ at the first observation time $\tau(x)$. In the same way we can rewrite the cost functional for (Y_t) giving

$$\lim_{T\to\infty} \mathbb{E}_x^u \left(\frac{1}{T} \int_0^T \tilde{C}(Y_s, u(Y_s))\, ds \right) = \sum_{k=1}^{d} \tilde{q}_k(x)\eta_k$$

where the weights $\tilde{q}_k(x) := \mathbb{P}_x^u(Y_t \in C_k \text{ for some } t > 0)$ are characterized by

$$\tilde{q}_k(x) = \sum_{y\in C_k} \frac{G_u(x,y)}{-G_u(x,x)} + \sum_{z\in D, z\neq x} \frac{G_u(x,z)}{-G_u(x,x)} \tilde{q}_k(z)$$

which is the law of total probability with respect to the state of (Y_t) after its first jump. We define the vector $p_k \in \mathbb{R}^{|D|}$ by $p_k(x) := \sum_{y\in C_k} \frac{G_u(x,y)}{-G_u(x,x)}$ and the matrix $A \in \mathbb{R}^{|D|,|D|}$ by $A(x,z) := \frac{G_u(x,z)}{-G_u(x,x)}$ and get

$$\tilde{q}_k = (I - A)^{-1} p_k$$

where $I \in \mathbb{R}^{|D|,|D|}$ is the identity matrix. On the other hand, we can convert (3.16) into

$$(1 - P_u(x,x)) q_k(x) = \sum_{y\in C_k} P_u(x,y) + \sum_{z\in D, z\neq x} P_u(x,z) q_k(z) \tag{3.17}$$

which - noting that by definition (3.10) it holds $P_u(x,x) = 1 + \tau(x) G_u(x,x)$ - is equivalent to

$$\begin{aligned} q_k(x) &= -\sum_{y\in C_k} \frac{P_u(x,y)}{\tau(x) G_u(x,x)} - \sum_{z\in D, z\neq x} \frac{P_u(x,z)}{\tau(x) G_u(x,x)} q_k(z) \\ &\overset{(3.10)}{=} -\sum_{y\in C_k} \frac{G_u(x,y)}{G_u(x,x)} - \sum_{z\in D, z\neq x} \frac{G_u(x,z)}{G_u(x,x)} q_k(z). \end{aligned}$$

In total we obtain

$$q_k = (I - A)^{-1} p_k = \tilde{q}_k$$

and with it the equality of average costs of the processes (X_t) and (Y_t).
This analysis can easily be extended to infinite lag times by observing that each state $x \in \mathcal{S}$ with $\tau(x) = \infty$ builds on its own a communication class $C_k = \{x\}$ for both processes $(X_{t_j})_{j \in \mathbb{N}_0}$ and $(Y_t)_{t \geq 0}$.[1] Then, for x with $\tau(x) = \infty$ we have

$$\begin{aligned}\lim_{T\to\infty} \mathbb{E}_x^u \left(\frac{1}{T} \int_0^T \tilde{C}(Y_s, u(Y_s))\, ds \right) &= \lim_{T\to\infty} \mathbb{E}_x^u \left(\frac{1}{T} \int_0^T \tilde{C}(x, a(x), \infty)\, ds \right) \\ &= \tilde{C}(x, a(x), \infty) \\ &= J(x,u),\end{aligned}$$

compare (3.13) and (2.2), while for the states with finite lag times the average costs are as before given by the weighted average of constants η_k with conform weights for both processes. □

Example 3.2 (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 a simple 2-state-example with $\mathcal{S} = \{x_1, x_2\}$ and $\mathcal{A} = \mathcal{A}(x_1) = \mathcal{A}(x_2) = \{a_1, a_2\}$ 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 $c_{\mathcal{S}}(x_1) = 0$, $c_{\mathcal{S}}(x_2) = 10$, $c_{\mathcal{A}}(a_1) = 0$, $c_{\mathcal{A}}(a_2) = 2$. 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 $c_{\mathcal{S}}(x_2) = 10$. Depending on the application, these states could - e.g. within medical therapy - refer to "healthy" and "diseased" with costs representing health damage; or - within machine maintenance - refer to "efficient" and "broken" with costs representing loss of profit. 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. For the policy $u(x) = (a(x), \tau(x))$ we choose $a(x_1) = a_1$, $\tau(x_1) = 5$, $a(x_2) = a_2$, $\tau(x_2) = 2$. Calculating the corresponding transition matrix P_u, the stationary distributions μ and $\tilde{\mu}$ and the cost function $\tilde{C}$ delivers a constant of long-term average costs

$$\eta_u = \sum_{x \in \mathcal{S}} \tilde{\mu}(x) \tilde{C}(x, u(x)) = 1.5989.$$

Figure 2 shows the accumulated costs

$$J_t(x_1, u) := \sum_{j=0}^{n(t)-1} \left(\int_{t_j}^{t_{j+1}} c(X_s, a(X_{t_j}))\, ds + k_{\text{info}} \right) + \int_{t_{n(t)}}^{t} c(X_s, a(X_{t_{n(t)}}))\, ds, \tag{3.18}$$

$$n(t) := \max\{j \in \mathbb{N} : t_j \leq t\},$$

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

[1] Note that now the split-up of space for the underlying process $(X_t)_{t \geq 0}$ might disagree.

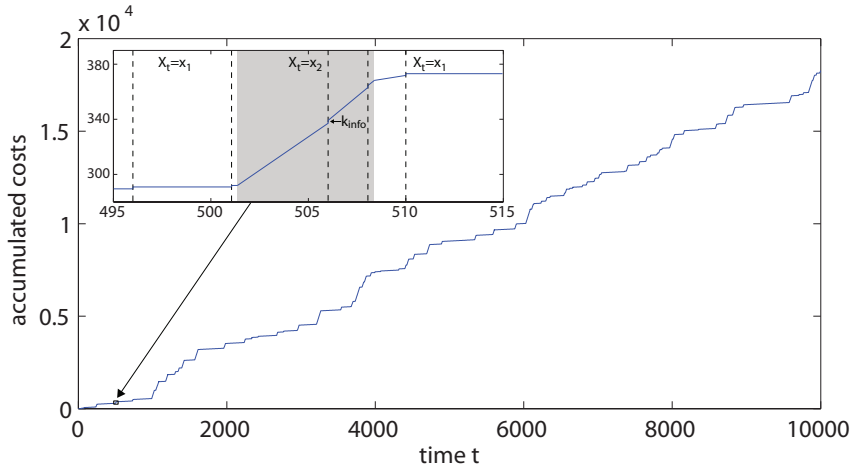

Figure 2: **Accumulated costs for $(X_t)_{t\geq 0}$.** Accumulated costs $J_t(x_1,u)$ defined in (3.18) up to time $t\geq 0$ for a trajectory of the MDP $(X_t)_{t\geq 0}$ with information costs k_{info} described in Example 3.2. The long-term asymptotics are given by $\eta_u \cdot t = 1.5989 \cdot t$. 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_{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.

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 $k_{\text{info}}=1$.
Equivalent graphs are given for the freely observable process $(Y_t)_{t\geq 0}$: Figure 3 shows the accumulated costs

$$\tilde{J}_t(x_1,u) := \int_0^t \tilde{C}(Y_s,u(Y_s))\,ds \tag{3.19}$$

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_{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_{info}) when the process is in state x_2.

The preceding analysis permits a straightforward translation of the original Markov control theory to the new setting of Markov control with information costs and policies with finite lag times. We just have to understand the results presented in a wide range of MDP literature in terms of the designed freely observable MDP $(Y_t)_{t\geq 0}$. This will be done now.

The optimal control problem consists of finding an average optimal policy. In order to guaranty the existence of an average-cost optimal policy, it is common standard to make several

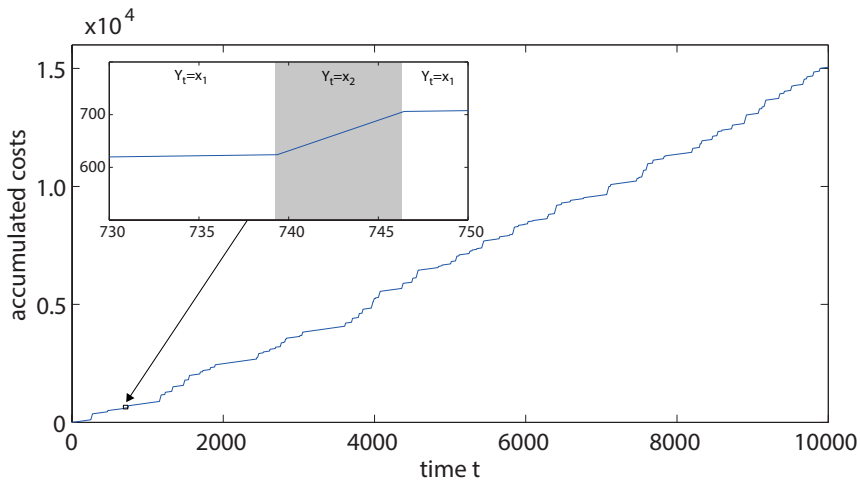

Figure 3: **Accumulated costs for the information process $(Y_t)_{t\geq 0}$.** Accumulated costs $\tilde{J}_t(x_1,u)$ defined in (3.19) up to time $t\geq 0$ for a trajectory of the equivalent freely observable MDP $(Y_t)_{t\geq 0}$ given Example 3.2. Like for the process $(X_t)_{t\geq 0}$ the long-term 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 $(Y_t)_{t\geq 0}$ 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.

assumptions concerning the transition rates and the other model parameters, compare [13, 15, 16, 35, 36] and others. Some of these assumptions become redundant in the case of a finite state space which is considered here. The relevant assumptions concern - on the one hand - the compactness of the action space and - on the other hand - the ergodicity of the process.

One of the relevant assumption is typically called *optimality condition* and states that the set of available actions $\mathcal{A}(x)$ is compact for each state x and that the cost function and the generator entries are continuous with respect to the action parameter. For the process $(Y_t)_{t\geq 0}$ the set of available actions for a state $x\in\mathcal{S}$ is of the form $\mathcal{A}(x)\times(0,\infty]$ which is not compact. However, we can state that, by $k_{\text{info}}>0$, it holds

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

such that we can locate a lower bound $\varepsilon>0$ for the optimal lag times. This way, the relevant set of actions can be restricted to the set $\mathcal{A}(x)\times[\varepsilon,\infty]$ and so the compactness condition is naturally fulfilled as long as $\mathcal{A}(x)$ is compact. It remains to note that the new cost function $\tilde{C}$ defined in (2.6) and the generators $G_{a,\tau}$ defined in (3.10) are all continuous in τ (compare the statement in (3.12) and the definition in (3.13) for $\tau=\infty$), such that the continuity condition only needs to be checked for $\mathcal{A}(x)$.

Assumption 3.3. a) The set of available actions $\mathcal{A}(x)$ is compact for each state $x\in\mathcal{S}$.
b) For all $x,y\in\mathcal{S}$ and $\tau\in(0,\infty]$, the functions $\tilde{C}(x,a,\tau)$ and $G_{a,\tau}(x,y)$ are continuous in $a\in\mathcal{A}(x)$.

The second relevant assumption concerns the ergodicity of the controlled process which - in the setting of finite state spaces and continuous dynamics - means that the process has to be irreducible. As for finite lag times, the process $(Y_t)_{t\geq 0}$ is irreducible as long as the observation process $(X_{t_j})_{j\in\mathbb{N}_0}$ is irreducible (compare again the definition of the generators $G_{a,\tau}$ defined in (3.10)) which leads to the following assumption.

Assumption 3.4. For each state $x\in\mathcal{S}$ it holds

$$e^{L_a\tau}(x,y)>0$$

for all $y\in\mathcal{S}$, $a\in\mathcal{A}(x)$ and $\tau\in(0,\infty)$.

The case of infinite lag times will require a separate investigation as it breaches the ergodicity condition by $G_{a,\infty}(x,y)=0\ \forall x,y\in\mathcal{S}$, see (3.11).

3.2 The average-cost optimality equation in the setting of information costs

Given the analogy of the processes $(X_t)_{t\geq 0}$ (MDP with information costs) and the corresponding information process $(Y_t)_{t\geq 0}$ (completely observable MDP), we will now reproduce some of the common results of original Markov control theory to the new setting of information costs.

Average costs η_u for a given policy $u\in\mathcal{U}_{\text{info}}$.

In the original theory, the constant of average costs for a (not necessarily optimal) policy $u\in\mathcal{U}$ may be characterized by a system of linear equations. The same holds within the setting of information costs:

Lemma 3.5. *Suppose that Assumption 3.4 holds and let $u\in\mathcal{U}_{\text{info}}$ be a given policy with finite lag times. Then we have the following facts.*

a) There exists a function $v:\mathcal{S}\to\mathbb{R}$ such that the corresponding constant η_u of long-run expected average costs defined in (2.8) *fulfills the equation*

$$\eta_u = \tilde{C}(x,a(x),\tau(x))+\sum_{y\in\mathcal{S}} G_{a(x),\tau(x)}(x,y)v(y) \quad \forall x\in\mathcal{S} \tag{3.20}$$

with $\tilde{C}$ resp. $G_{a,\tau}$ defined in (2.6) *resp.* (3.10).

b) The constant η_u is uniquely determined by (3.20) *and coincides with the first entry of the vector*

$$(E-G_u)^{-1}\tilde{C}_u,$$

where

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

and $\tilde{C}_u(x)=\tilde{C}(x,a(x),\tau(x))$.

The proof of Lemma 3.5 can be found in the Appendix.

In the case of *infinite* lag times the long-term average costs might depend on the initial state: Imagine a policy $u \in \mathcal{U}_{\text{info}}$ 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

$$J(x,u) \neq J(y,u).$$

This is consistent with the choice of the generator $G_{a,\infty}$, compare (3.11): 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 weighted average of $J(x,u)$ and $J(y,u)$.

We can conclude that the statement of Lemma 3.5 has 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 $J(x,u) \neq J(y,u)$ naturally excludes the referring policy u from being optimal as it either holds $J(x,u) > J(y,u)$ or $J(x,u) < J(y,u)$. 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

$$J(x, \tilde{u}) = C(x, a(y), \infty) \overset{(*)}{=} C(y, a(y), \infty) = J(y,u) < J(x,u).$$

The second equality $(*)$ is due to the fact that the underlying 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 $J(x,u) < J(y,u)$ 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. We will now use this insight for the analysis of optimal policies.

Optimal policy and value function.

Also the value function of optimal average costs is characterized by a system of equations, which is the central statement of Markov control theory. It is expressed in the Bellman equation which we will now formulate for MDP‘s with information costs.

Theorem 3.6 (Average cost optimality equation for a rarely observable MDP)**.** *Given the Markov control model* (2.1) *with information cost parameter* $k_{\text{info}} > 0$ *and nonnegative cost function* c*, consider the average cost criterion* $J(x,u)$*. Suppose that Assumptions 3.3 and 3.4 are satisfied. Then the following statements hold.*

a) There exists a function $v^* : \mathcal{S} \to \mathbb{R}$ *and a constant* $g \geq 0$ *satisfying*

$$g = \inf_{\substack{a \in \mathcal{A} \\ \tau \in (0,\infty]}} \left\{ C(x,a,\tau) + \frac{k_{\text{info}}}{\tau} + \sum_{y \in \mathcal{S}} G_{a,\tau}(x,y) v^*(y) \right\} \quad \forall x \in \mathcal{S}. \tag{3.21}$$

b) It holds $g = \inf_{u \in \mathcal{U}_{\text{info}}} J(x,u)$ *for all* $x \in \mathcal{S}$.

c) Any policy $u \in \mathcal{U}_{\text{info}}$ *realizing the minimum in* (3.21) *is average cost optimal.*

Proof. For the original theory of a completely observable Markov decision process these statements are well known, see e.g. Theorem 7.8 in [14] or Theorem 4.1 in [35]. I.e., they hold for the information process $(Y_t)_{t \geq 0}$ which - by construction and making use of Theorem 3.1 - directly implies a) and b). Part c) follows from Lemma 3.5. □

As described in Section 3.1, the existence of an optimal policy realizing the minimum in equation (3.21) is guaranteed by Assumption 3.3. Theorem 3.6 states that this optimal policy is deterministic and stationary. The value function of optimal average costs is given by a constant

$$\eta^* := \inf_{u \in \mathcal{U}_{\text{info}}} \eta_u = V(x) \quad \forall x \in \mathcal{S}. \tag{3.22}$$

At the same time, the "new" optimality equation (3.21) allows to transfer the common dynamic programming algorithms to the given framework of costly state observations. The numerical implementation requires a discretization with respect to the lag time parameter τ.

Example 3.2 (cont.) *The optimal policy for the 2-state-example 3.2 with different cost parameters is given in Table 1. 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). Increasing the state costs* $c_{\mathcal{S}}(x_2)$ *results in decreasing lag times but increasing average costs (compare row 1 and row 4). For very high information costs, testing cannot be afforded at all and action* a_2 *is applied for both states (see row 5).*

$c_{\mathcal{S}}(x_2)$	$c_{\mathcal{A}}(a_2)$	k_{info}	$a^*(x_1)$	$a^*(x_2)$	$\tau^*(x_1)$	$\tau^*(x_2)$	η^*
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
20	2	1	a_1	a_2	3.7	1.0	2.69
10	2	1000	a_2	a_2	∞	∞	7.00

Table 1: **Parameter dependent optimal policy and optimal average costs for the 2-state-example.** This table shows the optimal policy for Example 3.2 and different values of $c_{\mathcal{S}}(x_2)$, $c_{\mathcal{A}}(a_2)$ and k_{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 η^* (independent of the state).

4 Cost analysis

Given the Markov control model with information costs (2.1), an evident question is how the corresponding value function of optimal long-run average costs is related to the value function of the original control problem (without information costs). 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_{info}, see Definition 2.1. In order to reveal the connection between the two settings, we will - in a first step - analyze how the constant of average costs can be split up into different components. Furthermore, we will investigate the influence of the information cost parameter with respect to the value function and the optimal policy. We will complete the cost analysis by providing a short insight into the effect that deviations from the optimal policy might have regarding the related costs.

4.1 Splitting of Average Costs into Components

In the following, we intend to calculate - given a control policy - the part of the constant of average costs which is caused by the information costs. In the same time, this delivers a constant of *net costs* containing only the state and action costs.

Again, we assume ergodic dynamics and finite lag times. For the setting of information costs, the constant η_u of long-run average costs for a policy u is the $\tilde{\mu}$-weighted average of the cost function $\tilde{C}$ defined in (2.6), compare Lemma 2.3. By this definition it is obvious to split up the constant η_u of total average costs into *average information costs* η_{info} and *average net costs* $\eta_{\text{net}} = \eta_u - \eta_{\text{info}}$ by setting

$$\eta_{\text{info}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot \frac{k_{\text{info}}}{\tau(x)} = \frac{k_{\text{info}}}{\sum_{y\in\mathcal{S}} \mu(y)\tau(y)}, \qquad \eta_{\text{net}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot C(x,a(x),\tau(x)) \tag{4.23}$$

with C defined in (2.5).

If the cost function c is of the form $c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a)$ (which means that state and action costs are independent of each other) we can write

$$\tilde{C}(x,a,\tau) = C_{\mathcal{S}}(x,a,\tau) + c_{\mathcal{A}}(a) + \frac{k_{\text{info}}}{\tau},$$

where

$$C_{\mathcal{S}}(x,a,\tau) := \mathbb{E}_x^a \left(\frac{1}{\tau} \int_0^\tau c_{\mathcal{S}}(X_s)\, ds \right) \tag{4.24}$$

are the expected average state costs during the following time interval of constant control after starting in state x. This leads to a further splitting of the net costs $\eta_{\text{net}} = \eta_{\mathcal{S}} + \eta_{\mathcal{A}}$ into *average state costs* and *average action costs* by setting

$$\eta_{\mathcal{S}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot C_{\mathcal{S}}(x,a(x),\tau(x)), \qquad \eta_{\mathcal{A}} := \sum_{x\in\mathcal{S}} \tilde{\mu}(x) \cdot c_{\mathcal{A}}(a(x)). \tag{4.25}$$

In total, we get a split-up of the long-run average costs given a policy u into components of information costs, state costs and action cost

$$\eta_u = \eta_{\text{info}} + \eta_{\mathcal{S}} + \eta_{\mathcal{A}}.$$

We can now reformulate Lemma 3.5, which characterizes the constant η_u of total average costs, for the components η_{info}, $\eta_{\mathcal{S}}$ and $\eta_{\mathcal{A}}$.

Lemma 4.1 (Cost splitting)**.** *Suppose that Assumption 3.4 holds and let $u \in \mathcal{U}_{\text{info}}$ be a given policy with finite lag times. Then the constant η_u of long-term average costs is given by a split-up of the form $\eta_u = \eta_{\text{info}} + \eta_{\mathcal{S}} + \eta_{\mathcal{A}}$, where the average information costs η_{info} are given by the first entry of the vector*

$$k_{\text{info}} \cdot (E - G_u)^{-1} \begin{pmatrix} \vdots \\ \frac{1}{\tau(x)} \\ \vdots \end{pmatrix},$$

while the average action costs $\eta_{\mathcal{A}}$ 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 $x \in \mathcal{S}$, and the average state costs $\eta_{\mathcal{S}}$ 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 (4.24)*.*

All three statements of Lemma 4.1 follow directly from Lemma 3.5 by setting for each case two of the three components $c_{\mathcal{S}}$, $c_{\mathcal{A}}$ and k_{info} in $\tilde{C}$ to zero, which results in a cost functional containing only information-, action- or state costs, respectively.

In the case of infinite lag times the average 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 average action costs of a state $x \in \mathcal{S}$ with infinite lag time $\tau(x) = \infty$ are given by $c_{\mathcal{A}}(a(x))$ and the corresponding average state costs are given by $\lim_{T \to \infty} \mathbb{E}_x^{a(x)} \left(\frac{1}{T} \int_0^T c_{\mathcal{S}}(X_s)\, ds \right)$.

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 an application in medical science (compare [8, 33, 34]) the state costs can 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.

Comparison to the original Markov control problem.

In this setting of ergodic dynamics, assume that there exists an optimal policy $u^* \in \mathcal{U}_{\text{info}}$ and let

$$\eta^* = \eta_{u^*} = \eta^*_{\text{info}} + \eta^*_{\mathcal{S}} + \eta^*_{\mathcal{A}}$$

denote the split-up for the corresponding constant of optimal average costs according to the definitions given in (4.23) and (4.25). Naturally, the value η^* should exceed the optimal costs of the related original control problem where information is free of charge. A comparison of both settings is given in Figure 4 for the 2-state-example 3.2. Here, even the net costs $\eta^*_{\text{net}} = \eta^*_{\mathcal{S}} + \eta^*_{\mathcal{A}}$ (blue and green area) exceed the total costs of the original model. This relation holds in general

and should be intuitive: 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.

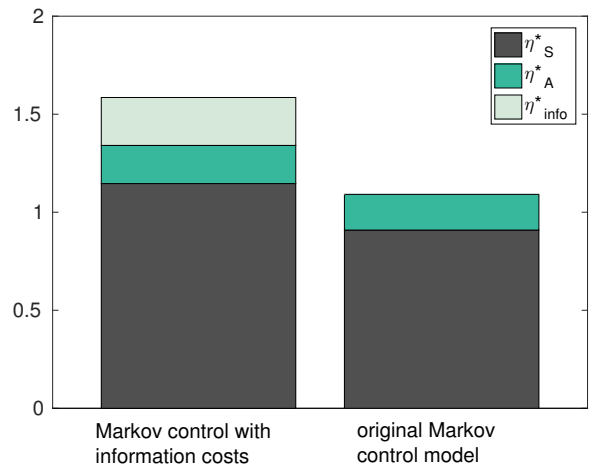

Figure 4: **Cost splitting for the 2-state-example.** The constant of optimal average costs η^* defined in (3.22) for Example 3.2 is divided into its components $\eta^*_{\mathcal{S}}$, $\eta^*_{\mathcal{A}}$ and η^*_{info} and compared to the free information case where the splitting is given by $\eta^*_{\mathcal{S}}+\eta^*_{\mathcal{A}}$. 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 1.

4.2 Monotonicity and Continuity with respect to k_{info}

Given the insight into the characteristics of the constant η^* and its different components, we now turn to the analysis of the central parameter k_{info} and its influence on the optimal costs resp. the optimal lag times. The questions to be answered in the following are: Is the constant of optimal average costs η^* monotone and continuous with respect to k_{info}? And how do the optimal lag times depend on k_{info}? Of special interest is the limit case $k_{\text{info}}\to 0$: Our observations will again confirm the intuitive relation to the original control problem with $k_{\text{info}}=0$. All statements are based on the ergodicity Assumption 3.4.

Monotonicity and continuity of η^* with respect to k_{info}.

Theorem 4.2 (Monotonicity of η^* with respect to k_{info})**.** *Let $\eta^*=\inf_{u\in\mathcal{U}_{\text{info}}} J(x,u)$ be the constant of optimal average costs for a given control problem with information costs k_{info}. Changing the parameter k_{info} of information costs 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. Let u^* be the optimal policy with respect to the parameter k_{info}, i.e. it holds $\eta^*=J(x,u^*)$ for all x. For a fixed policy, the cost functional J defined in (2.2), is obviously monotone in

k_{info}. This yields

$$\tilde{\eta}^* \leq \tilde{J}(x,u^*) \leq J(x,u^*) = \eta^*$$

where $\tilde{J}$ is the cost functional for the parameter $\tilde{k}_{\text{info}}$. □

As for the continuity of η^* with respect to k_{info} we have to distinguish between $k_{\text{info}} > 0$ and $k_{\text{info}} = 0$.

Lemma 4.3 (Continuity of η^* with respect to k_{info} at $k_{\text{info}} > 0$)**.** *The constant η^* of optimal average costs is continuous with respect to $k_{\text{info}} > 0$.*

Proof. We show continuity from the right. (For continuity from the left the argumentation is analogue.) For a given $k_{\text{info}} > 0$ consider the corresponding optimal policy u^* and the constant $\eta^* = J(x,u^*)$ of optimal average costs. Applying the policy u^* to the situation of higher information costs $\tilde{k}_{\text{info}} = k_{\text{info}} + \delta > k_{\text{info}}$, $\delta > 0$, increases only the information costs (as a part of the cost functional J), 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 $\tilde{J}$ - it holds

$$\tilde{J}(x,u^*) \leq \frac{k_{\text{info}} + \delta}{k_{\text{info}}} \eta^* \quad \forall x \in \mathcal{S}.$$

Let $\tilde{\eta}^*$ denote the optimal average costs given the parameter $\tilde{k}_{\text{info}}$. From $\tilde{\eta}^* \leq \tilde{J}(x,u^*) \, \forall x \in \mathcal{S}$ it follows

$$\tilde{\eta}^* \leq \frac{k_{\text{info}} + \delta}{k_{\text{info}}} \eta^* = \eta^* + \frac{\delta}{k_{\text{info}}} \eta^*,$$

and, using the monotonicity of η^* with respect to k_{info} (see Lemma 4.2),

$$0 < \tilde{\eta}^* - \eta^* \leq \frac{\delta}{k_{\text{info}}} \eta^*.$$

Hence, given an $\varepsilon > 0$, we can choose $\delta < \frac{\varepsilon \cdot k_{\text{info}}}{\eta^*}$ to guarantee $|\tilde{\eta}^* - \eta^*| < \varepsilon$, which completes the proof. □

Lemma 4.4 (Continuity of η^* with respect to k_{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_{info}. Let η_0 denote the optimal average costs of the corresponding original Markov Control model (without information costs). It holds*

$$\eta^*_{k_{\text{info}}} \xrightarrow{k_{\text{info}} \to 0} \eta_0.$$

Proof. For the proof we need the insight that Lemma 3.5 can be applied to the original case of vanishing information costs $k_{\text{info}} = 0$ where a policy only consists of choosing an action $a(x)$ for each state: One has to replace $G_{a(x),\tau(x)}$ by $L_{a(x)}$ and $\tilde{C}(x,a(x),\tau(x))$ by $c(x,a(x))$.
Now let $a_0 : \mathcal{S} \to \mathcal{A}$ be the optimal policy of the original Markov control problem. Defining $c_{a_0}(x) := c(x,a_0(x))$ and $L_{a_0}(x,y) := L_{a_0(x)}(x,y)$ for $x,y \in \mathcal{S}$, the application of Lemma 3.5 delivers

$$\eta_0 = c_{a_0}(x) + (L_{a_0} v)(x) \quad \forall x \in \mathcal{S},$$

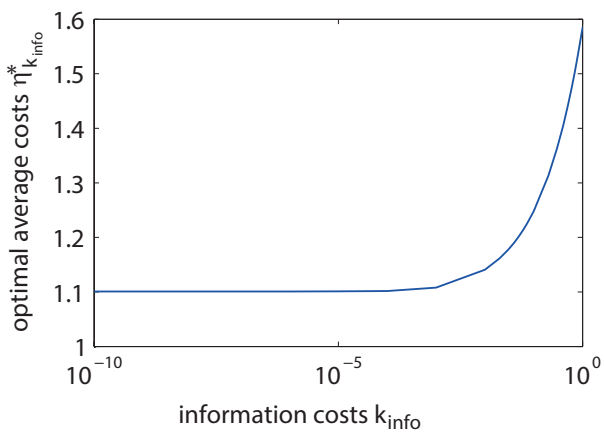

Figure 5: **k_{info} vs. optimal average costs for the 2-state-example.** The constant of optimal average costs $\eta^*_{k_{\text{info}}}$ for Example 3.2 depending on the information costs k_{info} in a logarithmic scale. For $k_{\text{info}}=1$ the values agree with the one given in Table 1 ($\eta^*=1.59$). For $k_{\text{info}}=10^{-10}$ the value is close to the optimal costs of the original model ($\eta^*\approx 1.09$).

where 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 3.5 the average costs $\eta_{k_{\text{info}}}$ induced by this policy u are given by the first component of

$$v_{k_{\text{info}}} := (E-G_u)^{-1}\tilde{C}_u,$$

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 $\frac{k_{\text{info}}}{\tau^*}=\frac{k_{\text{info}}}{\sqrt{k_{\text{info}}}}\to 0$. By the definition of G_u given in (3.10) and the properties of the generator matrix L_{a_0} it holds $G_u \xrightarrow{\tau^*\to 0} L_{a_0}$. Next, using $c_a(x):=c(x,a)$ and the properties of the generator matrices L_a, we have

$$\mathbb{E}_x^u\left(\frac{1}{\tau^*}\int_0^{\tau^*} c(X_s,a_0(x))\,ds\right)=\frac{1}{\tau^*}\int_0^{\tau^*}\left(e^{L_{a_0(x)}s}c_{a_0(x)}\right)(x)\,ds \xrightarrow{\tau^*\to 0} c_{a_0(x)}(x).$$

Putting everything together we get

$$v_{k_{\text{info}}} \xrightarrow{k_{\text{info}}\to 0} (E-L_{a_0})^{-1}c_{a_0},$$

which, again applying Lemma 3.5 to the case $k_{\text{info}}=0$, delivers η_0 in its first component, i.e. it holds $v_{k_{\text{info}}}(1)\xrightarrow{k_{\text{info}}\to 0}\eta_0$. Noting that $\eta_0\le\eta^*_{k_{\text{info}}}\le v_{k_{\text{info}}}(1)=\eta_{k_{\text{info}}}$ completes the proof. □

Figure 5 shows the optimal average costs η^* as a function of k_{info} for the 2-state-example 3.2.

Connection between k_{info} and τ^*.

Knowing that the constant η^* of optimal average costs is monotone and continuous in k_{info}, we now analyze the structure of the optimal policy depending on k_{info}. How do the optimal lag times $\tau^*(x)$ change when k_{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 4.5 (No monotonicity in τ^*)**.** *We consider a 3-state-model with $\mathcal{S}=\{x_1,x_I,x_2\}$ where x_I refers to an "intermediate" state, $\mathcal{A}=\{a_1,a_2\}$ and*

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

as well as $c(x,a)=c_{\mathcal{S}}(x)+c_{\mathcal{A}}(a)$ with $c_{\mathcal{S}}(x_1)=c_{\mathcal{S}}(x_I)=0$, $c_{\mathcal{S}}(x_2)=10$, $c_{\mathcal{A}}(a_1)=0$, $c_{\mathcal{A}}(a_2)=2$. Similar to Example 3.2, 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_{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_{info}. For $k_{\text{info}}<0.12$ the optimal action is given by $a^*(x_I)=a_1$, while for $k_{\text{info}}\geq 0.12$ it holds $a^*(x_I)=a_2$. For x_1 and x_2 the optimal lag time τ^* is monotone and continuous in k_{info}. However, for the intermediate state x_I there is a point of discontinuity at $k_{\text{info}}=0.12$: 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_{info}, see Figure 6. The corresponding optimal average costs are shown in Figure 7.

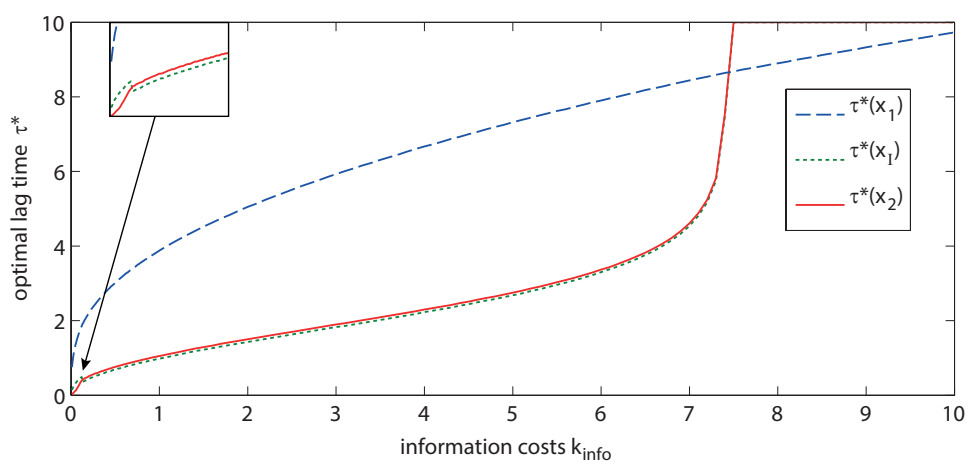

Figure 6: **k_{info} vs. optimal lag times for the 3-state-example 4.5.** The optimal lag times $\tau^*(x)$ for Example 4.5 depending on the information costs k_{info}. At $k_{\text{info}}=0.12$ the optimal lag time $\tau^*(x_I)$ of the intermediate state performs a jump.

Interpretation: This structure can be explained by the effect that both actions have on the process in the intermediate 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 in order to prevent the process from spending much time in state x_2. If the test is too expensive ($k_{\text{info}}>0.12$) such a safeguarding is not affordable and it is better to choose the "safe" (but more expensive) action a_2 in order to push the process towards the "good" state x_1. 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.
In order to understand what happens at the breaking point $k_{\text{info}}=0.12$ one can formulate "conditional" optimal policies: fixing the action for the intermediate state x_I to a_1 resp. a_2, the (conditional) optimal lag times $\tau_1^*(x_I)$ and $\tau_2^*(x_I)$ are continuous and monotone functions of k_{info}. The corresponding functions of conditional optimal costs have an intersection point at $k_{\text{info}}=0.12$. More precisely it holds $\eta_1^*(k_{\text{info}})-\eta_2^*(k_{\text{info}})<0$ (resp. $=0$ resp. >0) for $k_{\text{info}}<0.12$ (resp. $k_{\text{info}}=0.12$ resp. $k_{\text{info}}>0.12$). Now the connection should be clear: The overall minimal costs are the minimum of the conditional optimal costs.
For vanishing information costs the optimal lag times converge to zero, which gives a reasonable connection to the original Markov control model.

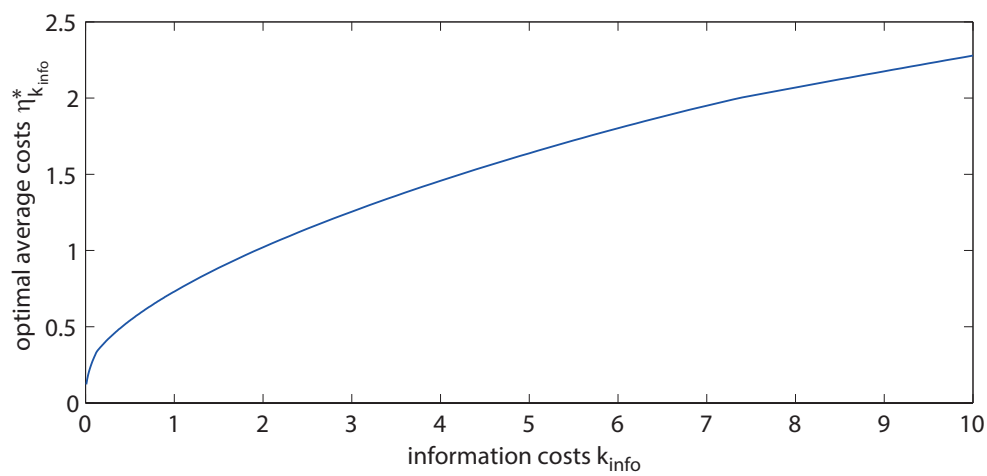

Figure 7: **k_{info} vs. optimal average costs for the 3-state-example 4.5.** The constant of optimal average costs for Example 4.5 depending on the information costs k_{info}, compare Lemma 4.4.

4.3 Sensitivity with respect to Lag Times $\tau(x)$

Finally, we shortly investigate the sensitivity of the constant of average costs with respect to the lag times τ: Given the optimal policy, how do small deviations from the optimal lag time τ^* affect the cost functional? The motivation is given by real-world applications: Due to practical restrictions, an exact adherence of the calculated optimal lag times might not be possible.

In fact, the answer to this question is already given in Section 2: The continuity of both the generators G_u and the cost functions $\tilde{C}_u$ with respect to $\tau(x)$ for all $x \in \mathcal{S}$ implies that the average costs η_u are continuous with respect to the lag time parameter, as well. 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 order to get an idea of how η_u depends on τ, we consider the 2-state-example 3.2. Figure 8 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 η_{net} are monotonously increasing in τ, whereas the total costs η exhibit a unique minimum. However, around these minima, there exists a wide area of values where the constant η of long-term average costs is nearly constant, which indicates a low sensitivity with respect to the lag times for this concrete example.

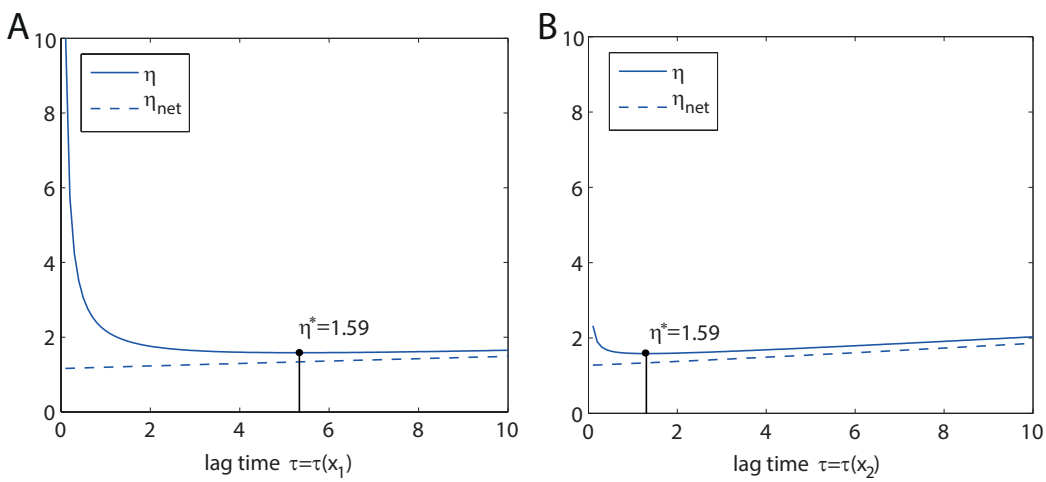

Figure 8: **Sensitivity with respect to τ for the 2-state-example.**
A: Lag time $\tau(x_1)$ vs. long-term average costs η and η_{net} for fixed $\tau(x_2)=1.3$ and $a(x_1)=1$, $a(x_2)=2$. B: Lag time $\tau(x_2)$ vs. long-term average costs η and η_{net} for fixed $\tau(x_1)=5.3$ and $a(x_1)=1$, $a(x_2)=2$.
All other parameters coincide with those given in the first line of Table 1, 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 η 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 1.

Remark 4.6. *As for numerical consequences we can state the following: In the case of a low sensitivity of the cost functional with respect to the lag time parameter τ – as it is given in panel A of Figure 8 – the problem of finding the minimum solution is numerically ill-conditioned. A gradient descent with respect to τ would be extremely slow because the gradient would almost vanish within a wide area around the minimum solution. However, 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...) and tests cannot be placed arbitrarily exact in time anyway, which overcomes this difficulty of exact optimization. For example, for medical*

tests or machine checkups a day declaration might be realistic and sufficient, compare the medical application proposed in [8, 33, 34]. This suggests to discretize the domain of the parameter τ in a suitable way. Furthermore, for practical purpose, a nearby solution is completely satisfying as long as the resulting costs are close to optimal, which is exactly given in this situation of low sensitivity.

5 Conclusion

We presented a quite general setting for continuous-time Markov decision processes which are not permanently observable. The observation of the process takes place at singular points in time which themselves are subject to the control of the decision maker. As every observation produces a fixed amount of information costs, a careful choice of the observation times is required in order to avoid an expansion of costs. Depending on the state observation, the decision maker chooses an action which determines the stochastic dynamics of the process within the next time interval of blind progress.

The approach is motivated by the fact that in many real-world applications a permanent observation and control of the process under consideration is not feasible. Especially situations in the context of medical therapy suggest that state examinations are costly and therefore rare, see [8, 33, 34] where the theory is applied in order to determine optimal treatment strategies against HIV-1. While these works consider the criterion of discounted costs, we here developed the theory for the criterion of long term average costs which requires a completely different approach. It turned out that the given MDP with incomplete information can be reformulated by an equivalent fully observable MDP. This main result allowed for an elegant transfer of the well-known theory to the new setting. Especially, we managed to reconstruct the Bellman equation which - as in the usual Markov control theory - characterizes both the optimal policy and the value function of optimal average costs.

Our analysis 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 further generalize the control model. Finding an answer to this question proves to be a topic of future research.

Acknowledgement

This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114.

APPENDIX

Proof of Lemma 3.5. 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

$$\rho = \tilde{C}(x, a(x), \tau(x)) + \sum_{y \in \mathcal{S}} G_{a(x),\tau(x)}(x,y) v(y) \quad \text{for all } x \in \mathcal{S}. \tag{5.26}$$

Due to the structure of the generator matrix G_u it holds $G_u(v+d) = G_u v$ for any constant vector $d \in \mathbb{R}^{|\mathcal{S}|}$ such that we can set $v(1) = \rho$ without loss of generality. Equation (5.26) can

now be written as $Ev = \tilde{C}_u + G_u v$ which yields

$$v = (E - G_u)^{-1} \tilde{C}_u.$$

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

$$(E - G_u) w = 0 \tag{5.27}$$

would have a solution $w \neq 0$. Now, equation (5.27) is equivalent to $w(1) = (G_u w)(x)$ for all $x \in \mathcal{S}$. However, for $x_{\min} := \arg\min w(x)$ and $x_{\max} := \arg\max w(x)$ we have

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

and

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

This leads to $0 \leq (G_u w)(x_{\min}) = w(1) = (G_u w)(x_{\max}) \leq 0$ which means that $w(1)$ has to be zero, and thus $G_u w = 0$ 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 ρ and v are uniquely defined by G_u and $\tilde{C}_u$. (Without this side constraint, the constant ρ 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 G_u, we consider the function $f : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ given by $f(c) = \left((E - G_u)^{-1} c\right)(1) = c(x) + (G_u (E - G_u)^{-1} c)(x) = c(x) + (G_u v_c)(x)$, 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 ρ depending on the cost function c. As f is obviously linear in c, by the Riesz representation theorem (see e.g. [30]) 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 $G_u c$ yields $f(G_u c) = (G_u c)(x) + (G_u (E - G_u)^{-1} c)(x) = (G_u c)(x) + (G_u v_c)(x) = G_u (c + v_c)(x) = \rho$ for all x. Now the last equality corresponds to equation (5.26) by setting the cost function to zero such that the constant ρ is given by $f(0) = \langle w, 0 \rangle = 0$. We get $f(G_u c) = 0$ and with it

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

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

References

[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] R. F. Anderson and A. Friedman. Optimal inspections in a stochastic control problem with costly observations ii. *Math. Operations Res.*, 3, 1978.

[3] J. Bather. An optimal stopping problem with costly information. *Bulletin of Institute for International Statistics*, 45:9–24, 1973.

[4] R. Bellman. A Markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957.

[5] R. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

[6] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

[7] B. T. Doshi. Continuous time control of Markov processes on arbitrary state space: average return criterion. *Stochastic Processes and their Applications*, 4:55–77, 1976.

[8] S. Duwal, S. Winkelmann, C. Schütte, and M. von Kleist. Optimal treatment strategies in the context of 'treatment for prevention' against HIV-1 in resource-poor settings. *PLOS Computational Biology*, submitted, 2015.

[9] E. A. Feinberg and A. Shwartz. *Handbook of Markov Decision Processes.* Kluwer Academic Publishers, 2002.

[10] W. H. Fleming and R. W. Rishel. Deterministic and stochastic optimal control. In *Applications of Mathematics.* Springer, Heidelberg, 1975.

[11] X. P. Guo and O. Hernández-Lerma. Continuous-time controlled Markov chains. *The Annals of Applied Probability*, 13:363–388, 2003.

[12] X. P. Guo and O. Hernández-Lerma. Continuous-time controlled Markov chains with discounted rewards. *Acta Applicandae Mathematica*, 79:195–216, 2003.

[13] 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.

[14] X. P. Guo and O. Hernández-Lerma. Continuous-time Markov decision processes: theory and applications. In *Stochastic Modelling and Applied Probability.* Springer, Heidelberg, 2009.

[15] X. P. Guo and U. Rieder. Average optimality for continuous-time Markov decision processes in polish spaces. *The Annals of Applied Probability*, 16:730–756, 2006.

[16] X. P. Guo and W. P. Zhu. Denumerable state continuous time Markov decision processes with unbounded cost and transition rates under average criterion. *The ANZIAM Journal*, 43(04):541–557, 2002.

[17] 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.

[18] O. Hernández-Lerma and J. B. Lasserre. *Discrete-time Markov control processes: basic optimality criteria.* Springer, 1996.

[19] R. A. Howard. *Dynamic programming and Markov processes.* MIT Press, 1960.

[20] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[21] M. J. Kim. *Optimal Control and Estimation of Stochastic Systems with Costly Partial Information.* PhD thesis, University of Toronto, 2012.

[22] W. S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28:47–66, 1991.

[23] B. L. Miller. Finite state continuous time Markov decision processes with an infinite planning horizon. *Journal of mathematical analysis and applications*, 22:552–569, 1968.

[24] G. E. Monahan. A survey of partially observable Markov decision processes: theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

[25] James R Norris. *Markov chains.* Number 2008. Cambridge university press, 1998.

[26] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., 1994.

[27] M. Robin. Long-term average cost control problems for continuous time Markov processes: a survey. *Acta Applicandae Mathematicae*, 1:281–299, 1983.

[28] I. R. Savage. Surveillance problems. *Naval Research Logistics*, 9:187–209, 1962.

[29] E. J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Operations Research*, 26:282–304, 1978.

[30] D. Werner. *Funktionalanalysis.* Springer, 2005.

[31] D. J. White. Dynamic programming, Markov chains, and the method of successive approximations. *Math. Anal. Appl.*, 6:373–376, 1963.

[32] D. J. White. A survey of applications of Markov decision processes. *J Opl Res Soc*, 44:1073–1096, 1993.

[33] S. Winkelmann, C. Schütte, and M. 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*, 2(1), 2012.

[34] S. Winkelmann, C. Schütte, and M. von Kleist. Markov control processes with rare state observation: Theory and application to treatment scheduling in HIV-1. *Comm. Math. Sci.*, 12:859–877, 2014.

[35] 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.

[36] 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.