A Markov Decision Model for a Surveillance Application and Risk-Sensitive Markov Decision Processes

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des

Karlsruher Instituts für Technologie

genehmigte

DISSERTATION

von

Dipl.-Math. techn. Jonathan Theodor Ott

aus Karlsruhe

Tag der mündlichen Prüfung: 3. November 2010

Referentin: Prof. Dr. Nicole Bäuerle
Korreferent: Priv.-Doz. Dr. Dieter Kadelka

Vorwort

Diese Arbeit entstand während meiner nun fast dreijährigen Zeit als wissenschaftlicher Mitarbeiter am Institut für Stochastik des Karlsruher Instituts für Technologie. Sie wurde im Rahmen des von Frau Prof. Dr. Nicole Bäuerle geleiteten Projekts „Risikogesteuerte Umfeldexploration (REX)“ angefertigt, das vom Bundesministerium für Bildung und Forschung der Bundesrepublik Deutschland unter dem Förderkennzeichen 03BAPAC1 gefördert wurde.

Als Erstes gilt mein Dank meiner Betreuerin Frau Prof. Dr. Nicole Bäuerle für ihr stetes Interesse an den Problemen und Ergebnissen meiner Arbeit sowie für sich daraus ergebende fachliche Diskussionen. Außerdem möchte ich Herrn Priv.-Doz. Dr. Dieter Kadelka für die Übernahme des Korreferats sowie für die inspirierende Zusammenarbeit danken.

Herrn Dr. Jürgen Geißler und Frau Dr. Elisabeth Peynsipp-Byma sowie den weiteren Mitarbeitern der Abteilung Interaktive Analyse des Fraunhofer IOSB möchte ich für die fruchtbare Zusammenarbeit, die Beiträge und die regen Diskussionen, die sich im Laufe von REX ergaben, danken. Insbesondere möchte ich Frau Dipl.-Inform. Jutta Hild für ihre Mitarbeit und ihre Begeisterung für die mathematischen Probleme von REX danken. Zudem möchte ich Herrn Dr. Peter Klausmann und Herrn M. Sc. Dipl.-Inform. (FH) Ralph Majer der Vitracom AG für die Beteiligung an REX danken.

Weiter möchte ich mich bei allen Mitarbeitern des Instituts für Stochastik für das angenehme Arbeitsklima bedanken, unter dem sich die letzten drei Jahre anregend forschen ließ.

Mein großer Dank gehört meiner Großfamilie für das Vertrauen, das sie in mich gesetzt hat, diese Dissertation fertigzustellen, sowie für deren Unterstützung und ständige Motivation, dieses Ziel zu erreichen. Zuletzt danke ich meinen Freunden.

Karlsruhe, im September 2010 Jonathan Ott

Contents

List of Figures

1. Introduction

Critical infrastructures are the backbone of every industrialized society. Therefore, they demand extraordinary protection. Multiple threats such as accidents, natural hazards, crime and terrorism menace these infrastructures. The main goal of the security staff of such a critical infrastructure is to protect the infrastructure from disastrous events. To this end, the security staff has to perform the best response action according to the current threat situation.

In order to achieve this goal, surveillance of the critical infrastructure is the first choice. Several sensors gather information on the current threat situation of the infrastructure. These data are evaluated by the security staff or by computer algorithms. After the data have been evaluated, the security staff has to assess the current threat situation correctly. Thereafter, the security staff performs appropriate actions based on the risk assessment. In the following, closed infrastructures such as railway stations, airports and logistic centres are considered.

In state-of-the-art surveillance of critical infrastructures, the infrastructures are equipped with monitoring cameras and alarm-triggering sensors such as smoke detectors and photoelectric sensors. The decision support for the security staff consists of simple operation guidelines in form of a small book. These guidelines list in very detail the appropriate actions that should be executed when a certain alarm is triggered or when suspect observations are made from the cameras or other available surveillance means. The guidelines are based on an extensive risk assessment of the infrastructure and on experience of experts.

A trivial way to protect a critical infrastructure is to increase the number of sensors and to increase the size of the security staff so that the infrastructure is under complete control at any point in time. But this approach is very expensive and therefore not practicable. Increasing the number of sensors only, would lead to informational overflow since all information has to be evaluated by the rather small number of security staff. By this, the view on the essential point might be obscured. Therefore, decision support is desirable so that attention is focussed on the essentials. In this text, a mathematical model is presented from which decision support can easily be provided. The background of the present work is that the sensors and the security staff should be deployed in some kind of an intelligent manner in order to provide insight to the current threat situation or in order to prevent arising threats in the most efficient way. Here, "intelligence" comes into play by using the framework of finite-state finite-action continuous-time Markov decision processes (CMDPs). The solution of a CMDP provides optimal actions which take into account the evolution of the threat situations over the whole time horizon, which is assumed to be infinite in our case.

Mathematics finds use in numerous surveillance applications. But applications primarily focus on methods for sensor data analysis. One important matter is the evaluation of sensor data with statistical and simulation methods in order to discover knowledge from the data. This is known as data mining. Several methods are presented, e. g., in (Tan et al., 2006; Ganguly et al., 2009). Furthermore, biometrics, e. g., face, fingerprint and speech recognition, finds use in surveillance applications (cf. (Delac and Grgic, 2004; Tistarelli and Nixon, 2009; Tistarelli et al., 2009)), which also requires mathematical background in stochastics and statistics.

(Shu et al., 2005) present a framework for performing real-time event analysis. It consists of a smart surveillance system in which the sensor evaluation is automatic. From the result of the sensor evaluation, certain events of interest are automatically generated and cross indexed into a single data repository. The security staff then performs appropriate actions with the help of the database which also provides real-time access to recorded sensor data. This framework merges the data of the single surveillance means in order to give more information to the security staff. But the framework does not propose optimal actions to the security staff. Although there are systems which provide information on the current threat situation of infrastructures, no attempt can be found in the literature in order to support the security staff in choosing optimal actions. This thesis presents a first step in the direction of proposing optimal actions to the decision maker. To this end, a mathematical model for the dynamics of threat of a critical infrastructure is formalized, from which optimal actions can be obtained for the current threat situation.

In this work, we present the mathematical model for the dynamics of threat of a critical infrastructure. The resulting model is a CMDP. Several theoretical aspects of the model are discussed. Among these, the structure of optimal policies is considered and bounds on the optimal value function are derived. Unfortunately, the exact solution cannot be computed for large infrastructures due to the sheer complexity of the model. Therefore, we consider a new heuristic index, and we examine the mathematical background of this index for a generalization of the model for the surveillance task. The surveillance task itself is approximately solved by a heuristics which uses the aforementioned index. Furthermore, we show that the original framework might not be adequate to the decision maker's preferences since she might be risk-averse. Therefore, finite-state finite-action discrete-time Markov decision processes (MDPs) are considered with respect to new optimality criteria. At first, the criterion is to minimize the average value-at-risk of the total discounted cost over a finite

or over an infinite horizon. Then the minimization of the average value-at-risk of the average cost is considered. The thesis is organized as follows.

In chapter 2, we shortly present the mathematical theory of CMDPs. The theory is well-known and can be found, e. g., in (Bertsekas, 2001; Guo and Hernández-Lerma, 2009; Puterman, 2005). At first, the definition of a CMDP is given. In section 2.2, we introduce the expected total discounted cost criterion. In section 2.3, the expected total discounted cost criterion is solved. Furthermore, solution methods such as linear programming and value iteration are presented.

In chapter 3, the mathematical model for the surveillance task is given in detail. At first, the infrastructure is divided into sectors. To capture the structure of the infrastructure, we model a dependency structure of the sectors. Current threat of the infrastructure is measured by a finite number of threat levels. The model dynamics is based on defining proper threat events, which are endowed with certain features such as state transition functions, costs and occurrence rates. Moreover, elementary actions are modelled, which are endowed with transition mechanisms, costs and a time until completion. From the model, we derive a CMDP. In the course of this chapter, an example infrastructure is constructed which is considered in the following text.

In chapter 4, we consider mathematical aspects of the model. At first, it is shown that it is beneficial to have a large size of security staff in hand if the staff does not earn a fixed salary but is only paid for actual working time. Furthermore, if the parametrization of the mathematical model satisfies certain assumptions, then we derive that a policy which is of control threshold type is optimal. From a generalization of the mathematical model in section 4.2, we derive bounds on the optimal value function of the surveillance tasks in terms of the sector-wise optimal value functions.

In chapter 5, approximate solution methods and heuristics are studied. This is necessary since the exact solution methods do not work for large infrastructures due to the so-called curse of dimensionality. At first, requirements on approximation methods for the surveillance task are stated. It is shown that value iteration does not satisfy these requirements. In section 5.3, we consider approximate linear programming where we exploit the special structure of the mathematical model. In section 5.4, we consider an approach using heuristics. At first, the concept of index policies is explained by considering the well-known indices of Gittins and Whittle. In section 5.4.4, a new heuristic index is introduced. It is shown that it can be seen as an approximation of the Whittle index. The heuristic index and the Whittle index are compared in a small simulation study for a resource allocation problem. In section 5.4.6, the heuristics for the surveillance task based on the heuristic index is presented. The chapter is concluded with a study on the impact of the model parameters on the performance of the heuristics.

In chapter 6, we show that the concept of the expected total discounted cost might not fit the decision maker's preferences since she might be risk-averse. To this end, we investigate two risk measures, value-at-risk and average value-at-risk, with respect to their suitability for the surveillance task. It turns out that the average value-at-risk is an adequate risk measure for the surveillance task.

In chapter 7, MDPs are examined. At first, a technique, called uniformization, is illustrated to convert the CMDP-model for the surveillance task into an MDP which is equivalent for the expected total discounted cost criterion over an infinite horizon. The optimality criterion is the minimization of the average value-at-risk of the total discounted cost. In section 7.5, the finite-horizon case is studied. The analysis of this problem is tackled by solving an intermediate criterion at first. The intermediate criterion penalizes costs only if they exceed some given threshold. It is shown that there are deterministic optimal policies for the intermediate criterion, which might depend on the history. In section 7.5.2, the average value-at-risk criterion is solved. Again, it is shown that there exist deterministic optimal policies for this criterion. Furthermore, the analysis shows how optimal policies might be constructed. In section 7.6, the infinite horizon case is studied. Again, an intermediate criterion is examined at first. As in the finite-horizon case, costs exceeding a given threshold value are penalized by the amount of excess. We consider two approaches. At first, we use results from the finite-horizon case and examine the case when the time horizon converges towards infinity. In the second approach, we derive an optimality equation for the intermediate criterion. Furthermore, deterministic optimal policies are constructed. From the intermediate criterion, optimal policies for the average value-at-risk criterion can be determined.

In chapter 8, we again consider MDPs. But here, we focus on the long-run average cost. At first, we illustrate the uniformization technique for CMDPs which determines an equivalent MDP for the expected average cost criterion. In this chapter, the optimality criterion is to minimize the average value-at-risk of the average cost. To this end, the average value-at-risk is computed for finite random variables and Markov reward processes in sections 8.2 and 8.3. In section 8.6, unichain and weakly communicating MDPs are considered. For these MDP-classes, every policy which is optimal with respect to the expected average cost criterion is also optimal for the average value-at-risk criterion for the average cost. In section 8.7, we study the average value-at-risk criterion for general MDPs for stationary policies. To this end, the state space is partitioned into its strongly communicating classes and its transient states. We restrict ourselves to MDPs in which the strongly communicating classes consist of exactly one state. Then, an appropriate non-convex non-differentiable mathematical program is formulated the solution of which is the minimal average value-at-risk of the average cost. The chapter is concluded with a remark, which shows the structural difference between the expected average cost criterion and the average value-at-risk criterion for the average cost.

2. Continuous-Time Markov Decision Processes

In chapter 3, the dynamics of threat of an infrastructure is modelled in terms of a finite-state finite-action continuous-time Markov decision process (CMDP). In this chapter, we provide the theoretic foundation of the arising surveillance task. We give the definition of a CMDP. Furthermore, we state the main results of CMDP-theory with respect to the expected total discounted cost criterion. Besides, two solution methods are presented: value iteration and linear programming.

2.1. Definition of Continuous-Time Markov Decision Processes

In this section, we define finite-state finite-action CMDPs. Furthermore, we define randomized history-dependent and deterministic stationary policies, which play a crucial role in the following. CMDPs are considered, e. g., in (Bertsekas, 2001; Guo and Hernández-Lerma, 2009; Puterman, 2005). The main results and definitions of this chapter are obtained from (Puterman, 2005).

Definition 2.1.1. A *finite-state finite-action continuous-time Markov decision process (CMDP)* Γ is given by $\Gamma = (S, A, D, \Lambda, P, \mathscr{K}, \mathscr{C}, \alpha)$, where the components are defined as follows:

- S is a non-empty finite *state space.*
- A is a non-empty finite *action space.*
- $D \subset S \times A$ is the *restriction set,* which satisfies the condition $D(s) \coloneqq \{a \in A : (s,a) \in D\} \neq \emptyset$, $s \in S$. The set $D(s)$ contains all actions which are admissible in state $s \in S$.
- $\Lambda = (\lambda(s,a))_{(s,a)\in D}$, where $\lambda(s,a) \geq 0$, $(s,a) \in D$, are the *transition rates.*
- $P = (p^a_{ss'})_{s,s' \in S, a \in D(s)}$, where $p^a_{ss'} \geq 0$ and $\sum_{s' \in S} p^a_{ss'} = 1$, s, $s' \in S$, $a \in D(s)$, are the *transition probabilities.*
- $\mathscr{K} = (K(s,a))_{(s,a)\in D}$, where $K(s,a) \in \mathbb{R}$ is the *setup cost* when initiating action $a \in D(s)$ in state $s \in S$.
- $\mathscr{C} = (C(s,a))_{(s,a)\in D}$, where $C(s,a) \in \mathbb{R}$, is the *cost rate,* which has to be paid continuously as long as the state is $s \in S$ and action $a \in D(s)$ is executed.
- $\alpha > 0$ is a discount rate.

A CMDP develops over time as follows: at time $t_0 = 0$, the system occupies some initial state $s \in S$. The decision maker chooses an action a from the set $D(s)$, which could possibly be done by an appropriate random mechanism. According to this choice, the system state remains in s according to an $\mathrm{Exp}(\lambda(s,a))$-distributed time T_1. At time $t_1 = T_1$, the system state jumps to some state $s' \in S$, possibly $s' = s$, with probability $p^a_{ss'}$. The cost the decision maker has to pay for the time interval $[t_0, t_1)$ is given by the setup cost $K(s,a)$ and an additional cost $C(s,a) \cdot T_1$ given by the cost rate and the sojourn time. By choosing the next action $a' \in D(s')$, again, the system stays an $\mathrm{Exp}(\lambda(s',a'))$-distributed time T_2 in s' and jumps to another state $s'' \in S$ with probability $p^{a'}_{s's''}$ at time $t_2 = t_1 + T_2$. The cost for the time interval $[t_1, t_2)$ is given by $K(s',a') + C(s',a') \cdot T_1$. In this manner, the system state is evolving over time. A state path of a three-state and two-action CMDP is illustrated in Figure 2.1, where the actions are chosen randomly according to a uniform distribution and $t_k = \sum_{i=1}^{k} T_i$, $k = 0, 1, \ldots$, are the decision times.

Note that the decision maker is only allowed to choose actions at transition times, i. e., she must not abort a currently being executed action. In CMDPs, the decision maker has to decide which action is the most appropriate, i. e., optimal action in some sense. The basis of decision making is captured in a so-called history.

Definition 2.1.2. For $k \in \mathbb{N}_0$, we define the *set of histories (up to decision epoch k)* recursively by

$$H_0 \coloneqq S,$$
$$H_{k+1} \coloneqq H_k \times A \times [0, \infty) \times S.$$

An element of the set H_k is called a *history (up to decision epoch k).* For every $k \in \mathbb{N}_0$, a history $h_k = (x_0, a_0, t_0, x_1, \ldots, a_{k-1}, t_{k-1}, x_k)$ is called *admissible* if $a_l \in D(x_l)$ for all $0 \leq l \leq k-1$.

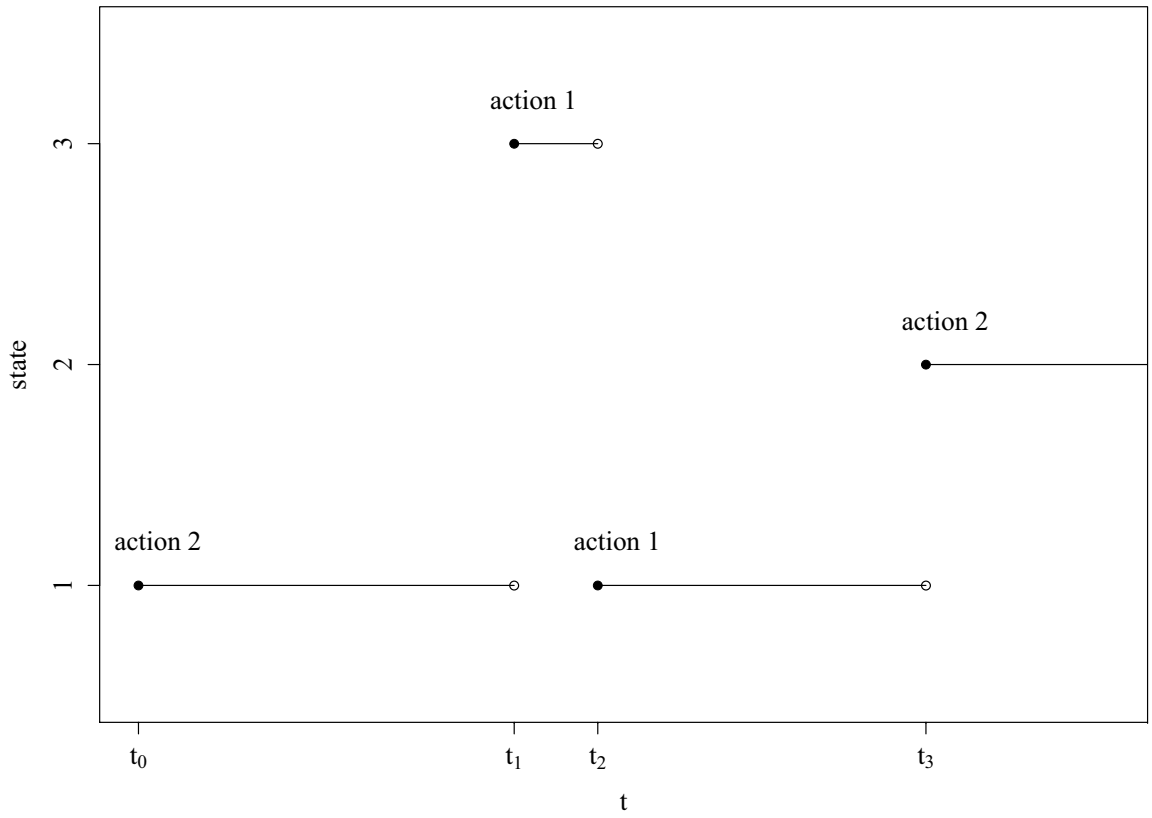


Figure 2.1.: State path of a CMDP.

The most general form of how the decision maker can choose an action when she has the knowledge of the evolution of the process within our framework is given in the next definition.

Definition 2.1.3. A *randomized history-dependent policy* $\pi = (\pi_k)_{k\in\mathbb{N}_0}$ is a sequence of measurable mappings $\pi_k : H_k \times A \to [0,1]$ such that

$$\sum_{a\in D(x_k)} \pi_k(h_k,a) = 1 \quad \text{and} \quad \pi_k(h_k,a) = 0 \quad (a \notin D(x_k))$$

for all $k \in \mathbb{N}_0$ and all admissible $h_k = (x_0,a_0,t_0,x_1,\ldots,a_{k-1},t_{k-1},x_k) \in H_k$. We also write $\pi_k(a|h_k) := \pi_k(h_k,a)$. Further, let Π be the *set of all randomized history-dependent policies.*

By this definition, the decision maker chooses an action randomly depending on the history. Another important class of policies is the class of deterministic stationary policies where the action choice depends on the current state only.

Definition 2.1.4. A policy $\pi \in \Pi$ is a *deterministic stationary policy* if $\pi_k(a|x_0,a_0,t_0,x_1,\ldots,a_{k-1},t_{k-1},x_k) = \pi_k(a|x'_0,a'_0, t'_0,x'_1,\ldots,a'_{k-1},t'_{k-1},x_k) =: \pi_k(a|x_k) \in \{0,1\}$ for all $x_l, x'_l \in S$, $a_l \in D(x_l)$, $a'_l \in D(x'_l)$ and for all $t_l, t'_l \in [0,\infty)$, $l = 0,\ldots,$ $k-1$, for all $k \in \mathbb{N}_0$, and if $\pi_k(a|x_k) = \pi_l(a|x_k)$, $a \in D(x_k)$, for all $k, l \in \mathbb{N}_0$. A deterministic stationary policy can be identified with a constant sequence $\mu^\infty := (\mu,\mu,\ldots)$ of some *decision rule* $\mu : S \to A$ such that $\mu(s) \in D(s)$ for all $s \in S$ so that μ^∞ can be identified with μ itself.

There are more classes of policies of interest, depending on whether the action choice is randomized or deterministic and whether the action choice is stationary or not. But in view of Theorem 2.3.4, these classes are of little interest for the surveillance application in the following chapters.

Next, similar to (Puterman, 2005), section 11.1.4, we construct a suitable probability space $(\Omega,\mathscr{A},P^\pi)$ for a given CMDP for every $\pi \in \Pi$. Let $\pi \in \Pi$. Then define

$$\Omega := \bigtimes_{k=0}^{\infty} (S \times A \times [0,\infty)) \quad \text{and} \quad \mathscr{A} := \bigotimes_{k=0}^{\infty} (\mathscr{P}(S) \times \mathscr{P}(A) \times \mathscr{B}([0,\infty)))$$

where $\mathscr{P}(M)$ is the power set of the set M and $\mathscr{B}([0,\infty))$ is the Borel σ-algebra of $[0,\infty)$. An element $\omega \in \Omega$ has the typical form

$$\omega = (x_0, a_0, \tau_0, x_1, a_1, \tau_1, \ldots).$$

Define the projection mappings $X_k(\omega) := x_k$, $A_k(\omega) := a_k$ and $T_k(\omega) := \tau_k$, $k \in \mathbb{N}_0$, as the state, action and sojourn time processes respectively. We assume that the initial state is known so that the initial probability P_0 is degenerated to a single state $s \in S$ with $P_0(\{s\}) = 1$. The policy π induces a unique probability measure P^π on $(\Omega, \mathscr{A})$ by setting the conditional probabilities

$$\begin{aligned}
&P^\pi(X_0 = x_0) = P_0(\{x_0\}),\\
&P^\pi(A_k = a_k \,|\, X_0 = x_0, A_0 = A_0, T_0 = \tau_0, \ldots, X_k = x_k) = \pi_k(a_k|(x_0, a_0, \tau_0, \ldots, x_k)),\\
&P^\pi(T_k \le \tau_k \,|\, X_0 = x_0, A_0 = a_0, T_0 = \tau_0, \ldots, X_k = x_k, A_k = a_k) = 1 - e^{-\lambda(x_k,a_k)\,\tau_k},\\
&P^\pi(X_{k+1} = x_{k+1} \,|\, X_0 = x_0, A_0 = a_0, T_0 = \tau_0, \ldots, X_k = x_k, A_k = a_k, T_k = \tau_k) = p^{a_k}_{x_k x_{k+1}},\\
&\qquad\qquad x_k \in S,\, a_k \in A,\, \tau_k \in [0,\infty),\, k = 0, 1, \ldots,
\end{aligned}$$

with regard to the Ionescu-Tulcea theorem (cf. (Bertsekas and Shreve, 1978), Proposition 7.28) since S, A and $[0,\infty)$ are Borel spaces. The expectation with respect to the probability measure P^π is denoted by E^π. The process of transition times $(t_k)_{k\in\mathbb{N}_0}$ is given by $t_k = \sum_{i=1}^{k} T_i$, $k \in \mathbb{N}_0$. The state process $(\tilde{X}_t)_{t\ge 0}$ and the action process $(\tilde{A}_t)_{t\ge 0}$ regarded as processes in continuous time are defined by

$$\tilde{X}_t := X_k \quad \text{and} \quad \tilde{A}_t = A_k, \quad \text{if } t \in [t_k, t_{k+1}) \text{ for some } k \in \mathbb{N}_0, \quad t \ge 0.$$

2.2. Expected Total Discounted Cost Criterion

Given a CMDP, one could consider several criteria such as the expected average cost criterion or the expected total discounted cost criterion. In the following, we consider the expected total discounted cost over an infinite time horizon. The *total discounted cost* C^∞ *under policy* π is defined by

$$C^\infty := \sum_{k=0}^{\infty} \left[e^{-\alpha t_k} K(X_k, A_k) + \int_{t_k}^{t_{k+1}} e^{-\alpha t} C(X_k, A_k)\, dt \right] = \sum_{k=0}^{\infty} e^{-\alpha t_k} K(X_k, A_k) + \int_0^\infty e^{-\alpha t} C(\tilde{X}_t, \tilde{A}_t)\, dt,$$

where the processes $(X_k)_{k\in\mathbb{N}_0}$, $(\tilde{X}_t)_{t\ge 0}$, $(A_k)_{k\in\mathbb{N}_0}$, $(\tilde{A})_{t\ge 0}$ and $(T_k)_{k\in\mathbb{N}_0}$ depend on the underlying $\pi \in \Pi$ and so does C^∞. Note that under every policy $\pi \in \Pi$, the total discounted cost C^∞ is a random variable. The setup cost is discounted at the corresponding decision time, whereas the cost rate between to decision times t_k and t_{k+1} is discounted continuously over the whole interval $[t_k, t_{k+1})$, $k = 0, 1, \ldots$. Denote the *expected total discounted cost under policy* $\pi \in \Pi$ when the initial state is $s \in S$ by

$$v^\pi(s) := E^\pi[C^\infty \,|\, X_0 = s].$$

If $\mu : S \to A$ is a decision rule, then we write $v^\mu := v^{\mu^\infty}$. The value function v^π is finite for every $\pi \in \Pi$ since D is finite. We seek for a policy that minimizes the expected total discounted cost when starting in some state $s \in S$. To this end, let the *optimal value function* v^* be defined by

$$v^*(s) := \inf_{\pi\in\Pi} v^\pi(s), \quad s \in S.$$

Definition 2.2.1. Let $\varepsilon > 0$. A policy $\pi \in \Pi$ is *ε-optimal (with respect to the expected total discounted cost criterion)* if $v^\pi \le v^* + \varepsilon$. A policy $\pi^* \in \Pi$ is *optimal (with respect to the expected total discounted cost criterion)* if $v^{\pi^*} = v^*$.

By definition of the infimum, an ε-optimal policy exists for every $\varepsilon > 0$. But the existence of an optimal policy is not assured a priori. Anyway, we shall see that a deterministic stationary optimal policy exists within the framework of a finite-state finite-action CMDP.

If the decision maker is not interested in costs but in rewards, the expected total discounted reward criterion, which is to find a policy $\pi^* \in \Pi$ such that $E^{\pi^*}[-C^\infty \,|\, X_0 = s] = \sup_{\pi\in\Pi} E^\pi[-C^\infty \,|\, X_0 = s]$, $s \in S$, is equivalent to the expected discounted cost criterion.

For a given initial state $x_0 \in S$, another standard criterion for CMDPs is to minimize the expected average cost

$$\limsup_{t\to\infty} \frac{1}{t} E^\pi \left[\int_0^t C(\tilde{X}_t, \tilde{A}_t)\, dt + \sum_{k=0}^{\nu_t - 1} K(X_k, A_k) \,\middle|\, X_0 = x_0 \right]$$

over all $\pi \in \Pi$, where ν_t, $t \geq 0$, is the number of state transitions that have occurred up to time t and $(\nu_t)_{t\geq 0}$ is the respective process. This criterion is also considered in (Puterman, 2005). Equivalently, the expected average reward criterion could be considered.

2.3. Solution Methods for the Expected Total Discounted Cost Criterion

In this section, we present several solution methods based on (Puterman, 2005). First, let us define the expected discounted one-step cost by

$$\begin{aligned} c(s,a) &:= E\left[K(s,a) + \int_0^{T_0} e^{-\alpha t} C(s,a)\, dt\right] = K(s,a) + \frac{C(s,a)}{\alpha} E\left[1 - e^{-\alpha T_0}\right] \\ &= K(s,a) + \frac{C(s,a)}{\alpha}\left[1 - \int_0^{\infty} \lambda(s,a)\, e^{-\lambda(s,a)t}\, e^{-\alpha t}\, dt\right] \\ &= K(s,a) + \frac{C(s,a)}{\alpha}\left[1 + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha}\left[e^{-(\lambda(s,a)+\alpha)t}\right]_0^{\infty}\right] = K(s,a) + \frac{C(s,a)}{\lambda(s,a)+\alpha}, \quad (s,a) \in D, \end{aligned}$$

where $T_0 \sim \mathrm{Exp}(\lambda(s,a))$. Then $c(s,a)$ is the expected discounted cost that occurs during the first decision time $t_0 = 0$ and the subsequent decision time t_1. For every function $v : S \to \mathbb{R}$, we define the one-step cost operator T by

$$Tv(s) := \min_{a \in D(s)} \left\{ c(s,a) + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p^a_{ss'}\, v(s') \right\}, \quad s \in S.$$

The next three lemmas state important properties of the operator T.

Lemma 2.3.1. *The operator T is isotonic, i. e., for u, $v : S \to \mathbb{R}$ such that $u \leq v$, we have $Tu \leq Tv$.*

Proof. This is immediately clear from the definition of T since $\lambda(s,a) \geq 0$ for all $(s,a) \in D$ and since $p^a_{ss'} \geq 0$ for all s, $s' \in S$, $a \in D(s)$. □

Lemma 2.3.2. *It holds $Tv^* = v^*$.*

Proof. After applying Theorem 11.1.1 of (Puterman, 2005), the proof is essentially given in (Puterman, 2005), Theorem 6.22. □

The proof of this lemma is quite involved because, among other aspects, one has to show that one can restrict attention to decision rules rather than to the general class of randomized history-dependent policies. Again, the details can be found in (Puterman, 2005).

Lemma 2.3.3. *The operator T is a contraction mapping on the function space $V := \{v \mid v : S \to \mathbb{R}\}$ with Lipschitz constant $\gamma^* := \max_{(s,a)\in D}\{\lambda(s,a)/(\lambda(s,a)+\alpha)\}$ with respect to the supremum norm.*

Proof. Let u, $v : S \to \mathbb{R}$ be two arbitrary functions. Further, let $\delta := \|u - v\|_\infty$ such that $u(s) - \delta \leq v(s) \leq u(s) + \delta$ for all $s \in S$. Then

$$\begin{aligned} T(u+\delta)(s) &= \min_{a \in D(s)} \left\{ c(s,a) + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p^a_{ss'}\, (u+\delta)(s') \right\} \\ &= \min_{a \in D(s)} \left\{ c(s,a) + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p^a_{ss'}\, u(s') + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha}\, \delta \right\} \leq Tu(s) + \delta \max_{a \in D(s)} \left\{ \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \right\} \\ &\leq Tu(s) + \delta\, \gamma^*, \quad s \in S. \end{aligned} \tag{2.1}$$

Analogously,

$$T(u-\delta)(s) \geq Tu(s) - \delta\, \gamma^*, \quad s \in S. \tag{2.2}$$

Since T is isotonic, we have

$$Tu(s) - \delta\, \gamma^* \overset{(2.2)}{\leq} T(u-\delta)(s) \leq Tv(s) \leq T(u+\delta)(s) \overset{(2.1)}{\leq} Tu(s) + \delta\, \gamma^* \quad (s \in S),$$

from which we conclude $\|Tu - Tv\|_\infty \leq \gamma^* \|u - v\|_\infty$. Since D is finite, we have $\gamma^* < 1$. □

Now, we present the main theorem for finite-state finite-action CMDPs. It shows that the optimal value function is the unique fixed point of the operator T, that minimizing actions are optimal in the respective states and that an optimal decision rule exists.

Theorem 2.3.4. *The optimal value function v^* is the unique solution to the optimality equation, also called Bellman equation,*

$$v = Tv \tag{2.3}$$

in the function space $V \coloneqq \{v \mid v : S \to \mathbb{R}\}$. Furthermore, a deterministic stationary policy $(\mu^*)^\infty$ is optimal if and only if

$$\mu^*(s) \in \operatorname*{arg\,min}_{a \in D(s)} \left\{ c(s,a) + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p_{ss'}^a v^*(s') \right\} \quad (s \in S) \tag{2.4}$$

if and only if

$$c(s,\mu^*(s)) + \frac{\lambda(s,\mu^*(s))}{\lambda(s,\mu^*(s))+\alpha} \sum_{s' \in S} p_{ss'}^{\mu^*(s)} v^*(s') = v^*(s) \quad (s \in S). \tag{2.5}$$

Furthermore, such a decision rule μ^* exists.

Proof. Since $(V, \|\cdot\|_\infty)$ is a complete vector space and T is a contraction mapping on V, the Banach fixed point theorem (cf. (Agarwal et al., 2009), Theorem 4.1.5) provides a unique solution $v' \in V$ to the equation $v = Tv$. By Lemma 2.3.2, $v' = v^*$. Now, let $\mu : S \to A$ such that $\mu(s) \in D(s)$, $s \in S$, be an arbitrary decision rule, then v^μ is the unique fixed point of the equation $v(s) = c(s,\mu(s)) + \lambda(s,\mu(s))/(\lambda(s,\mu(s))+\alpha) \sum_{s' \in S} p_{ss'}^{\mu(s)} v(s')$, $s \in S$, by setting $D(s) = \{\mu(s)\}$, $s \in S$, for the just derived result. Defining μ^* such that $\mu^*(s) \in \arg\min_{a \in D(s)} \left\{ c(s,a) + \lambda(s,a)/(\lambda(s,a)+\alpha) \sum_{s' \in S} p_{ss'}^a v^*(s') \right\}$, $s \in S$, we obtain

$$v^*(s) = Tv^*(s) = c(s,\mu^*(s)) + \frac{\lambda(s,\mu^*(s))}{\lambda(s,\mu^*(s))+\alpha} \sum_{s' \in S} p_{ss'}^{\mu^*(s)} v^*(s'), \quad s \in S,$$

and therefore, $v^* = v^{\mu^*}$ so that μ^* is optimal. There exists some $\mu^*(s) \in \arg\min_{a \in D(s)} \{c(s,a) + \lambda(s,a)/(\lambda(s,a)+\alpha) \cdot \sum_{s' \in S} p_{ss'}^a v^*(s')\}$ for every $s \in S$ since $D(s) \neq \emptyset$ is finite for all $s \in S$. Conversely, let $\mu : S \to A$ be an optimal decision rule. Since $v^\mu = v^*$, we have

$$v^*(s) = v^\mu(s) = \underbrace{c(s,\mu(s)) + \frac{\lambda(s,\mu(s))}{\lambda(s,\mu(s))+\alpha} \sum_{s' \in S} p_{ss'}^{\mu(s)} v^\mu(s')}_{=: T_\mu v^\mu(s)} \geq Tv^\mu(s) = Tv^*(s) = v^*(s), \quad s \in S. \tag{2.6}$$

Hence, equality holds in (2.6). Thus, we have $\mu^*(s) \in \arg\min_{a \in D(s)} \left\{ c(s,a) + \lambda(s,a)/(\lambda(s,a)+\alpha) \sum_{s' \in S} p_{ss'}^a v^*(s') \right\}$, $s \in S$, since otherwise, we had $v^\mu(s) = T_\mu v^\mu(s) = T_\mu v^*(s) > Tv^*(s) = v^*(s)$ for some $s \in S$, violating the optimality of μ. Equivalence of (2.4) and (2.5) is evident by applying $v^* = Tv^*$. □

Remark 2.3.5. For a given decision rule $\mu : S \to A$, the corresponding value function can be obtained by solving the system of linear equations

$$v(s) = c(s,\mu(s)) + \frac{\lambda(s,\mu(s))}{\lambda(s,\mu(s))+\alpha} \sum_{s' \in S} p_{ss'}^{\mu(s)} v(s'), \quad s \in S.$$

This system of equations has exactly one solution due to the proof of Theorem 2.3.4.

Theorem 2.3.4 justifies the terminology of an "optimal action a in state s" since there is an optimal decision rule. For applications, this result is important since one only has to search the set of decision rules in order to find an optimal policy. Having this result in hand, it essentially remains to solve the optimality equation (2.3). This can be done in several ways. We demonstrate two important methods in the following sections. Another common method is Howard's policy improvement algorithm often referred to as "policy iteration" (cf. (Puterman, 2005), section 6.4).

2.3.1. Value Iteration

Since the proof of Theorem 2.3.4 relies on the Banach fixed point theorem, a straightforward method for approximately solving CMDPs with expected total discounted cost criterion can be deduced as in (Puterman, 2005), Theorem 6.3.1, from the standard bounds given by the Banach fixed point theorem.

Proposition 2.3.6. *Let $v_0 : S \to \mathbb{R}$ be an arbitrary function. Then $T^n v_0 \to v^*$, $n \to \infty$. Let $\varepsilon > 0$. If $\|v^{n+1} - v^n\|_\infty < \varepsilon(1-\gamma^*)/(2\gamma^*)$, then a deterministic stationary policy μ^∞ such that*

$$\mu(s) \in \operatorname*{arg\,min}_{a \in D(s)} \left\{ c(s,a) + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p^a_{ss'} v^{n+1}(s') \right\}, \quad s \in S,$$

is ε-optimal.

Proof. See (Puterman, 2005), Theorem 6.3.1. □

This method is called *value iteration* since it is an iterated application of T to some initial (value) function. One drawback of this method is that it does not compute the exact optimal value function in general. Only an ε-approximation of the optimal value function can be established theoretically. Nonetheless, the ε-optimal decision rule μ might indeed be optimal. Also note that the function $c(s,a) + \lambda(s,a)/(\lambda(s,a)+\alpha)\sum_{s' \in S} p^a_{ss'} v^{n+1}(s')$ has to be evaluated for every $(s,a) \in D$ at each iteration step so that for very large state spaces or very large restriction sets respectively the value iteration method is inappropriate.

2.3.2. Linear Programming

In order to obtain the exact optimal value function and a deterministic stationary optimal policy, one can use linear programming. The next lemma provides the basis of this approach.

Lemma 2.3.7. *The optimal value function v^* is the largest function $v : S \to \mathbb{R}$ such that $v \leq Tv$.*

Proof. Let v^* be the optimal value function, i. e., v^* is the unique solution to the optimality equation $v = Tv$ by Theorem 2.3.4. Then we trivially have $v^* \leq Tv^*$. So, v^* is a solution to $v \leq Tv$. To show that v^* is the largest solution to $v \leq Tv$, let $\hat{v}$ be an arbitrary solution to $v \leq Tv$. We have to show that $v^* \geq \hat{v}$. This is clear since we have $\hat{v} \leq T\hat{v} \leq T^2\hat{v} \leq \ldots \leq T^k\hat{v} \overset{k\to\infty}{\to} v^*$ by exploiting that T is isotonic when using induction and by the value iteration method of Proposition 2.3.6. Hence, v^* is the largest solution to $v \leq Tv$. □

Indeed, this lemma is required to prove Lemma 2.3.2 so that it has to be proven without the use of the Bellman equation. Now, we can formulate a linear program the solution of which is the optimal value function. Moreover, we are able to obtain an optimal decision rule from its solution.

Theorem 2.3.8. *Let $\zeta : S \to \mathbb{R}_{>0}$ be arbitrary. The optimal value function v^* is the unique solution to the following linear program* (LP)*:*

$$\begin{aligned} &\textit{Maximize} \sum_{s \in S} \zeta(s)\, v(s) \\ &\textit{under the constraint} \\ &v(s) - \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p^a_{ss'} v(s') \leq c(s,a) \quad ((s,a) \in D). \end{aligned} \tag{LP}$$

For $v = v^*$, the constraint $(s,a^*) \in D$ is active if and only if a^* is optimal in s.

Proof. By Lemma 2.3.7, v^* is the largest solution to

$$v \leq Tv \quad \Leftrightarrow \quad v(s) - \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s' \in S} p^a_{ss'} v(s') \leq c(s,a) \quad ((s,a) \in D).$$

Suppose that $\hat{v} \neq v^*$ is a solution to $v \leq Tv$. Then we have $\sum_{s \in S} \zeta(s)\hat{v}(s) < \sum_{s \in S} \zeta(s) v^*(s)$ since $\zeta > 0$. Hence, v^* is the unique solution to (LP). For $v = v^*$, we have that the constraint corresponding to $(s,a^*) \in D$ is active if, by definition,

$$v^*(s) - \frac{\lambda(s,a^*)}{\lambda(s,a^*)+\alpha} \sum_{s' \in S} p^{a^*}_{ss'} v^*(s') = c(s,a^*),$$

which holds if and only if a^* is optimal in s by Theorem 2.3.4. Conversely, we know by Theorem 2.3.4 that there is a deterministic stationary policy so that for every $s \in S$ there is some $a^* \in D(s)$ such that $c(s,a^*) + \lambda(s,a^*)/(\lambda(s,a^*) + \alpha) \sum_{s' \in S} p_{ss'}^{a^*} v^*(s') = v^*(s)$. In other words, the constraint corresponding to (s,a^*) is active for $v = v^*$. □

Remark 2.3.9.

1. Since an optimal decision rule exists according to Theorem 2.3.4, the active constraints of (LP) are the basis of an optimal decision rule. The decision rule μ^* is optimal if and only if the constraint $(s, \mu^*(s))$ is active for $v = v^*$ for every $s \in S$.

2. In section 6.9, (Puterman, 2005) also considers linear programming for solving the expected total discounted cost criterion. But he uses the dual formulation of the above linear program, which is the following for an arbitrary $\zeta > 0$:

$$\begin{aligned} &\text{Minimize} \sum_{(s,a) \in D} c(s,a)\, x(s,a) \\ &\text{under the constraints} \\ &\sum_{a \in D(s')} x(s',a) - \frac{\lambda(s,a)}{\lambda(s,a) + \alpha} \sum_{(s,a) \in D} p_{ss'}^{a}\, x(s,a) = \zeta(s') \quad (s' \in S), \\ &x(s,a) \geq 0 \quad ((s,a) \in D). \end{aligned} \tag{LP$'$}$$

   The advantage from a theoretic point of view is that existence of stationary optimal policies can be established without use of the optimality equation. Let x^* be a solution to (LP$'$). Then a stationary policy π, which is randomized in general, such that $\pi(a|s) = x^*(s,a) / \sum_{a' \in D(s)} x^*(s,a')$, $(s,a) \in D$, is optimal where $\pi(a|s)$ is the probability of choosing action a in state s. But from a practical point of view, this approach has the disadvantage that the optimal value function cannot be derived from the solution to the dual linear program (LP$'$). But as we shall see later, we need the optimal value functions of given CMDPs. Moreover, we can easily establish an optimal decision rule from the solution to (LP).

3. The Surveillance Task

In this chapter, we model the dynamics of threat of a critical infrastructure that is exposed to multiple threats. The dynamics are formulated in terms of a CMDP. Then the associated so-called *surveillance task* consists in minimizing the expected total discounted cost of the arising CMDP. This chapter is structured as follows: at first, we give a short overview on applications of CMDPs. In the following sections, we motivate the use of a CMDP for the threat model and describe the components of the model in detail. The chapter is closed by a small consideration of aspects concerning the practical application of the model.

3.1. Literature Overview of Applications of CMDPs

CMDPs find wide use in applications. In this section, we shortly summarize some areas of application besides the standard area of queueing systems (cf. (Bertsekas, 2001; Guo and Hernández-Lerma, 2009; Puterman, 2005)).

(White, 1993) lists numerous applications of Markov decision processes in discrete time (MDPs) as well as in continuous time. For the definition of MDPs, we refer to section 7.2. Continuous-time models with an infinite horizon are used in the following articles:

(Crabhill, 1974) considers a production system of various machines which could break down. The decision maker is able to repair several machines at the same time. The criterion considered is the expected average cost criterion.

(Deshmukh and Winston, 1979) consider pricing strategies for a dominant firm in some industry with a number of competing firms of different sizes. The dominant firm has to find a price of a product so that the expected total discounted reward is maximized. Here, the main model assumption is that, when the price of the product is high, then new firms are founded which want to skim the market and, therefore, decrease the profit of the dominant firm. So, the dominant firm has to find a price which balances maximal short profit and the loss of sales due to the foundation of new firms.

(Lefèvre, 1981) considers the control of the spread of an infectious disease for a closed population, which is modelled as a controlled birth and death process. The decision maker is able to quarantine the entire population or to apply some medical care treatments to the infective individuals. The goal is to minimize the expected total discounted cost.

More recent applications of CMDPs can be found in the following articles: (Qiu and Pedram, 1999) consider dynamic power management. It is sought for a policy which minimizes the power dissipation of a system of electric power consumers under performance constraints. Their goal is to minimize the expected total discounted cost as well as the expected average cost where the cost contains power consumption and a switching energy which arises at state transitions.

(Nair and Bapna, 2001) consider internet service providers, which offer services to users. It is assumed that there are two classes of customers, platinum and gold customers, which have different arrival and service rates. Since the internet service provider has a limited capacity, it must be decided whether arriving customers should be admitted to the system or whether they should be refused. The goal is to maximize expected total discounted reward.

(Economou, 2003) considers immigration of batches of pests in a habitat. The state space is the number of the pests in the habitat which is countably infinite. The damage done by the pests yields certain costs. The decision maker can set a catastrophe mechanism which also yields a cost. The expected total discounted cost criterion as well as the expected average cost criterion are considered.

(Kyriakidis, 2006) considers a population of pests in a habitat of finite size. As long as the decision maker does not intervene, pests immigrate into the habitat at given rates. The decision maker is able to introduce a predator to the habitat. The predator arrives at a given rate. When the predator arrives, immigration of the pests immediately stops. The predator captures the pests one by one at given rates until there is no pest left. After some time, the predator emigrates from the habitat and the pests commence to immigrate. The expected average cost criterion is considered.

(Guo et al., 2009) model biochemical reactions. The state is given by the actual number of several molecules. The actions are given as the rates of chemical reactions, which can be controlled by experimental conditions such as temperature and pH-value. The actual reward is given by weight factors of the respective molecules. They consider the expected average reward criterion.

(Feng and Pang, 2010) consider a manufacturer who produces one product. The decision maker decides whether the produced product should be sold on the spot market and at which rate the product should be produced in presence of a long-term contract. The considered optimization criterion is the expected total discounted cost criterion.

3.2. Motivation for the Use of a CMDP

To give a motivation for the use of a CMDP to model the dynamics of threat, we have a look at an example threat scenario which could take place at some fictitious airport. The evolution of threat could be as follows. At first, we assume that each sector of the airport is in a perfect safe state. Then, after some random time, some security- or safety-relevant event occurs, e. g., an alarm occurs in the technical sector. Due to this event, an increase of threat of the technical sector would be observed. Moreover, threat might be increased in other sectors because there is a dependency of spatial or functional nature. For example, after an alarm occurred in the technical sector, it is more likely that the electric power supply breaks down, and therefore, threat of sectors which require electric power, e. g., such as the terminal and the apron, might be increased. Moreover, threat might be increased in the terminal if it is proximate to the technical sector. After the occurrence of the alarm in the technical sector, the question arises which appropriate actions the security personnel should perform in the particular sectors of the airport in order to provide the safest states of the infrastructure over an infinite horizon. Specifically, if the security staff is limited, one has to decide where the actions should be performed.

These dynamics can be modelled as a CMDP if we assume that the occurrence times of relevant events such as alarms are exponentially distribution. The preceding ideas are formalized in the subsequent sections 3.3.1–3.3.5.

3.3. Components of the Threat Model

In this section, we describe the components of the threat model from which the CMDP is built. In detail, the components of the model are

- the structure of the infrastructure (cf. section 3.3.1),
- threat levels (cf. section 3.3.2),
- threat events (cf. section 3.3.3),
- elementary actions (cf. section 3.3.4) and
- dangerous objects (cf. section 3.3.5).

Throughout this section, we make things clear on an accompanying example. We consider a small airport consisting of four sectors: technical sector (TS), terminal (T), apron (A) and fence (F).

3.3.1. The Structure of the Infrastructure

The infrastructure itself is identified with its sectors. A particular sector might be some physical part of the infrastructure, e. g., a building, a floor of a building or a single room. But also a more abstract but critical component of the infrastructure could be modelled as a sector, e. g., the power supply or the water supply of the infrastructure. In the following, let Σ be the non-empty finite set of all sectors of the considered infrastructure.

As seen in the motivation in section 3.2, the sectors alone do not cover the whole structure of the threat dynamics. Dependencies between sectors have to be considered. To model the dependencies, we define the adjacency matrix $N \in \{0,1\}^{\Sigma\times\Sigma}$, where $N(\sigma,\sigma^*) = 0$, σ, $\sigma^* \in \Sigma$, if σ^* does not depend on σ, i. e., relevant events in σ do not have any influence concerning the threat state of σ^*. If σ^* depends on σ, then we define $N(\sigma,\sigma^*) = 1$. For completeness, we define $N(\sigma,\sigma) = 0$ for all $\sigma \in \Sigma$. Note that the dependency structure need not be symmetric.

Example 3.3.1. In Figure 3.1, the fictitious example airport is illustrated. Here, $\Sigma = \{\mathrm{TS},\mathrm{T},\mathrm{A},\mathrm{F}\}$ and

$$N = \begin{array}{c} \begin{array}{cccc} \mathrm{TS} & \mathrm{T} & \mathrm{A} & \mathrm{F} \end{array} \\ \left(\begin{array}{cccc} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right) \end{array} \begin{array}{c} \mathrm{TS} \\ \mathrm{T} \\ \mathrm{A} \\ \mathrm{F} \end{array}.$$

In Figure 3.1, the dependencies between the sectors are indicated with arrows where an arrow leading from σ to σ^* if $N(\sigma,\sigma^*) = 1$.

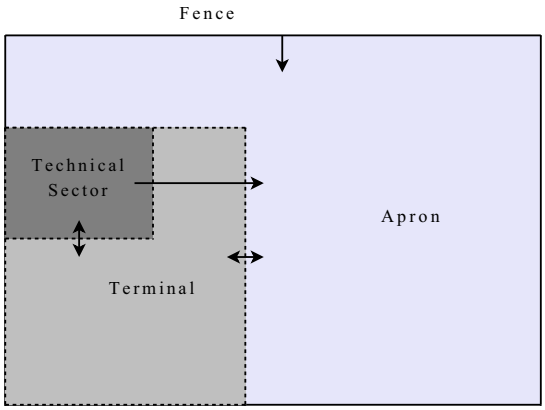


Figure 3.1.: Structure of the example airport.

3.3.2. Threat Levels

At every point in time, each sector is in a certain threat state. To measure the current threat state of a sector, we introduce *threat levels*. We assume that a threat level takes a value lying in the set $G = \{0, 1, \dots, g_{\max}\}$. Here, 0 should indicate a safe threat state, whereas $g_{\max}$ should indicate the most threatened state of a sector. An interpretation of the respective threat levels is given below since we have to define other components beforehand. However, we define

$$S \coloneqq G^{\Sigma}$$

as the state space of the CMDP. A state of the CMDP assigns a threat level to each sector. Therefore, the current state can easily be depicted as a risk map of the infrastructure from which the decision maker is able to determine the current threat state at a glance.

Example 3.3.2. In the airport example, we assume $g_{\max} = 4$ so that the state space of the respective CMDP is $S = \left\{ (s(\sigma))_{\sigma\in\Sigma} \,\middle|\, s(\sigma) \in \{0,1,2,3,4\},\ \sigma \in \Sigma \right\}$. The whole model consists of $5^4 = 625$ states.

3.3.3. Threat Events

We model events that influence threat of a certain sector with *threat events*. Examples for threat events of interest are various alarms in the sectors and the destruction of the sectors. The set of all threat events that may occur in sector $\sigma \in \Sigma$ are denoted by $\mathscr{E}(\sigma)$. Occurring threat events influence the threat situation in a crucial way. To this end, a threat event $e \in \mathscr{E}(\sigma)$ is endowed with certain features:

- When e occurs, the subsequent threat level of σ is determined by the transition function $\Psi_e : G \to G$.
- When e occurs, the subsequent threat levels of a dependent sector is determined by the transition function $\psi_e : G \to G$.
- When e occurs, the decision maker incurs the cost $C_e \geq 0$.
- If the current threat level of σ is g, then e will occur after an $\mathrm{Exp}(\lambda_e(g))$-distributed time which is independent of all occurrences of threat events of all sectors.

The cost C_e might be the actual monetary value of the threat event $e \in \mathscr{E}(\sigma)$. But C_e could also be a weight which represents the dislike of the decision maker for the occurrence of e.

By this parametrization, the subsequent state when $e \in \mathscr{E}(\sigma)$ occurred can be determined by a function $\zeta_e : S \to S$. If $s \in S$ is the current state just before $e \in \mathscr{E}(\sigma)$ occurs, then the subsequent threat level of $\sigma^* \in \Sigma$ is

$$\zeta_e(s)(\sigma^*) \coloneqq \begin{cases} \Psi_e(s(\sigma^*)), & \text{if } \sigma^* = \sigma \\ \psi_e(s(\sigma^*)), & \text{if } N(\sigma,\sigma^*) = 1 \\ s(\sigma^*), & \text{else} \end{cases}, \quad s \in S.$$

Example 3.3.3. In the airport example, we have three threat events for each sector. In every sector $\sigma \in \Sigma$, an alarm may be triggered in σ (e_σ^{alarm}), σ may be destroyed ($e_\sigma^{\text{destruction}}$), or nothing may have happened in σ ($e_\sigma^{\text{n. h.}}$). For every $\sigma \in \Sigma$, the transition functions are given by

$$\Psi_{e_\sigma^{\text{alarm}}}(g) = \min\{g+2, 4\}, \quad \psi_{e_\sigma^{\text{alarm}}}(g) = \min\{g+1, 4\},$$

$$\Psi_{e_\sigma^{\text{destruction}}}(g) = 0, \quad \psi_{e_\sigma^{\text{destruction}}}(g) = 4,$$
$$\Psi_{e_\sigma^{\text{n. h.}}}(g) = \begin{cases} 1, & \text{if } g = 0, 1, 2 \\ g-1, & \text{if } g = 3, 4 \end{cases}, \quad \psi_{e_\sigma^{\text{n. h.}}}(g) = g,$$

where $g \in G$. So, an alarm increases the threat level by two in the affected sector, and in dependent sectors, the threat level increases by one. When a sector is destructed, we assume that it is rebuilt in no time and that the subsequent threat level is 0. In contrast, all dependent sectors are severely threatened. We introduce a third threat event. It models that threat varies after some time has passed. The threat events $e_\sigma^{\text{n. h.}}$ are triggered by a random number generator which is assumed to be integrated in the system. When nothing has happened in a sector with threat levels 2, 3 or 4 during an appropriate amount of time, then the threat level decreases by one since there is evidence that threat is not as large as it is supposed to be. If the threat level is 0, then, after some time, threat increases since knowledge about the real situation diminishes. (This mechanism is introduced to force the decision maker to gather information or to initiate inspection walks after a random time. Otherwise, the decision maker would have to wait until an alarm is triggered or a sector is destructed. So, $e_\sigma^{\text{n. h.}}$ could be seen as a small alarm in σ.) If the current threat level is 1 and nothing happened, then the threat level remains 1. The costs of the threat events are

$$C_{e_\sigma^{\text{alarm}}} = 0 \quad (\sigma \in \Sigma),$$
$$C_{e_{\text{TS}}^{\text{destruction}}} = 1{,}000{,}000, \quad C_{e_{\text{T}}^{\text{destruction}}} = 10{,}000{,}000, \quad C_{e_{\text{A}}^{\text{destruction}}} = 5{,}000{,}000, \quad C_{e_{\text{F}}^{\text{destruction}}} = 100{,}000,$$
$$C_{e_\sigma^{\text{n. h.}}} = 0 \quad (\sigma \in \Sigma).$$

The occurrence rates of the threat events are given by

$$\lambda_{e_\sigma^{\text{alarm}}}(0) = \frac{1}{4380}, \quad \lambda_{e_\sigma^{\text{alarm}}}(1) = \frac{1}{84}, \quad \lambda_{e_\sigma^{\text{alarm}}}(2) = \frac{1}{12}, \quad \lambda_{e_\sigma^{\text{alarm}}}(3) = 2, \quad \lambda_{e_\sigma^{\text{alarm}}}(4) = 4,$$
$$\lambda_{e_\sigma^{\text{destruction}}}(0) = \frac{1}{8760}, \quad \lambda_{e_\sigma^{\text{destruction}}}(1) = \frac{1}{168}, \quad \lambda_{e_\sigma^{\text{destruction}}}(2) = \frac{1}{24}, \quad \lambda_{e_\sigma^{\text{destruction}}}(3) = 1, \quad \lambda_{e_\sigma^{\text{destruction}}}(4) = 2,$$
$$\lambda_{e_\sigma^{\text{n. h.}}}(0) = 1, \quad \lambda_{e_\sigma^{\text{n. h.}}}(1) = 1, \quad \lambda_{e_\sigma^{\text{n. h.}}}(2) = 1, \quad \lambda_{e_\sigma^{\text{n. h.}}}(3) = 1, \quad \lambda_{e_\sigma^{\text{n. h.}}}(4) = 1$$

for all $\sigma \in \Sigma$, where the time unit is given by 1 hour.

3.3.4. Elementary Actions

Now, we model the available actions of the CMDP. We assume that the decision maker is able to execute one *elementary action* from the finite set $A_0 \neq \emptyset$ in every sector. Further, we assume that a passive elementary action, denoted by 0, is available in every sector. Then the action space is

$$A := A_0^\Sigma,$$

i. e., the decision maker has to choose one elementary action for every sector. Similar to a threat event, an elementary action $a_0 \in A_0$ is endowed with certain features since it influences the dynamics of threat:

- There is a probabilistic transition mechanism $\Phi_{a_0,\sigma}$ for the affected sector σ. Let $g \in G$ be the current threat level of σ just before a_0 is completed. Then the subsequent threat level of σ is g' with probability $\Phi_{a_0,\sigma}^g(g')$.
- There is a transition function $\varphi_{a_0} : G \times G \times G \to G$ such that $\varphi_{a_0}(g_\sigma, g'_\sigma, g_{\sigma^*})$ is the subsequent threat level of the dependent sector σ^*, when due to the completion of elementary action a_0 in σ the respective threat level changes from g_σ to g'_σ and g_{σ^*} is the current threat level of σ^*.
- There is a cost rate $c_{a_0} \geq 0$ so that executing a_0 costs c_{a_0} per time unit.
- If a_0 is executed in σ, then it will be accomplished after an $\text{Exp}(\lambda_\sigma(a_0))$-distributed time which is independent of all occurrences of threat events of all sectors and independent of all completions of elementary actions in every sector.

For the passive elementary action 0, we define $\Phi_{0,\sigma}^g(g) = 1$ for all $\sigma \in \Sigma$ and $g \in G$, $\varphi_0(g_\sigma, g'_\sigma, g_{\sigma^*}) = g_{\sigma^*}$ for all g_σ, g'_σ, $g_{\sigma^*} \in G$, $c_0 = 0$ and for all $\lambda_\sigma(0) = 0$ for every $\sigma \in \Sigma$. So, the passive elementary action does not change the threat level of the respective sector, it does not cost anything, and it will never be accomplished. Therefore, a sector in which the passive action 0 is executed has to rely on elementary actions being executed in sectors from which the first one is

$g \setminus g'$	0	1	2	3	4
0	0.99	0	0	0	0.01
1	0.8	0	0	0	0.2
2	0.5	0	0	0	0.5
3	0.2	0	0	0	0.8
4	0.01	0	0	0	0.99

Table 3.1.: Parameters of $\Phi^g_{1,\sigma}(g')$ for all $\sigma \in \Sigma$.

dependent. If no such elementary action is executed, then the sector is at the mercy of threat events that could occur in the respective sector and in those sectors from which it is dependent.

Let $s = (s(\sigma))_{\sigma\in\Sigma} \in S$ be the current state just before completion of an elementary action a_0 in sector σ. Depending on the outcome $g' \in G$, the subsequent state can be determined by a function $\zeta^{g'}_{a_0,\sigma} : S \to S$ which is defined by

$$\zeta^{g'}_{a_0,\sigma}(s)(\sigma^*) := \begin{cases} g', & \text{if } \sigma^* = \sigma \\ \varphi_{a_0}(s(\sigma), g', s(\sigma^*)), & \text{if } N(\sigma,\sigma^*) = 1 \\ s(\sigma^*), & \text{else} \end{cases}, \quad s \in S,\ \sigma^* \in \Sigma.$$

Example 3.3.4. In the airport example, we have three elementary actions: "Do nothing" (0), "Camera evaluation" (1) and "Inspection walk" (2), where the parameters are the following: we give $\Phi^g_{1,\sigma}(g')$ in Table 3.1 which is independent of σ. When a camera evaluation is completed, it is observed that the sector is either in the safe threat level 0 or that it is severely threatened so that the threat level is 4. We define

$$\varphi_1(g_\sigma, g'_\sigma, g_{\sigma^*}) := \begin{cases} \min\left\{4, \max\left\{0, g_{\sigma^*} + \left\lfloor \frac{g'_\sigma - g_\sigma}{2} \right\rfloor\right\}\right\}, & \text{if } g_{\sigma^*} = 0 \\ \min\left\{4, \max\left\{1, g_{\sigma^*} + \left\lfloor \frac{g'_\sigma - g_\sigma}{2} \right\rfloor\right\}\right\}, & \text{else} \end{cases}, \quad g_\sigma, g'_\sigma \in G,$$

where $\lfloor\cdot\rfloor$ is the floor function, i. e., $\lfloor x \rfloor$ is the smallest integer z such that $x \le z$ for $x \in \mathbb{R}$. Thus, the subsequent threat level depends on the result of the camera evaluation in σ. So, by the definition of φ_1, we assume that we can conclude that the threat levels of dependent sectors change about a half of the change of the threat level of σ. The min-max construction assures that the threat level stays in $G = \{0,\dots,4\}$. The cost rate is $c_1 = 0$ since we assume that the security staff earns a fixed salary. The rates are

$$\lambda_\sigma(1) = 30 \quad (\sigma \in \Sigma),$$

which means that a camera evaluation takes approximately two minutes since the time unit is 1 hour. For the inspection walk, we assume

$$\Phi^g_{2,\sigma}(0) = 1 \quad (\sigma \in \Sigma,\ g \in G),$$

which means that the inspection walk is carried out perfectly leaving the respective sector at threat level 0. For the dependent sectors, we define the subsequent threat level by

$$\varphi_2(g_\sigma, g'_\sigma, g_{\sigma^*}) = \begin{cases} 0, & \text{if } g_{\sigma^*} = 0 \\ \max\{1, g_{\sigma^*} - 1\}, & \text{else} \end{cases}, \quad g_\sigma, g'_\sigma, g_{\sigma^*} \in G.$$

Again, we set the cost rate $c_2 = 0$. The rates until completion are given by

$$\lambda_\sigma(2) = 10 \quad (\sigma \in \Sigma)$$

so that an inspection walk takes about six minutes in every sector. The functions φ_1 and φ_2 are defined in such a way that it is not possible to reach threat level 0 by completing elementary actions in sectors from which the first one is dependent. To reach threat level 0, an elementary action has to be accomplished in the respective sector (or the respective sector must be destructed). This is due to the practitioner's point of view: when there is evidence that threat is increased in a certain sector, then it is necessary to act in that sector directly.

3.3.5. Dangerous Objects

In our model, we also include that there could be dangerous objects in the sectors which should be removed. By $\gamma_\sigma(g)$, we denote the probability that there is a dangerous object, e. g., a bomb or an abandoned suitcase, in sector σ at threat level g. The removal of the object costs an expected amount of $C_\sigma \geq 0$ which has to be paid when an elementary action is completed in the affected sector. In this manner, elementary actions are made more expensive.

Example 3.3.5. In the airport example, when elementary actions 1 or 2 are completed, there might be some probability of detecting dangerous objects. Nevertheless, we define $\gamma_\sigma(g) = 0$ for all $\sigma \in \Sigma$ and $g \in G$. We define $C_\sigma = 0$ for all $\sigma \in \Sigma$ so that removing dangerous objects is free of cost.

Now, we come back to the interpretation of a threat level $g \in G$ of sector $\sigma \in \Sigma$. The threat level is g if the threat events $e \in \mathscr{E}(\sigma)$ occur at the given rates $\lambda_e(g)$ and if dangerous objects are located in σ with probability $\gamma_\sigma(g)$.

3.3.6. Discount Rate

The discount rate $\alpha > 0$ determines the influence which the elapsed time has on the incurred cost. If α is very large, then only costs arising during a short period have a significant influence on the total discounted cost. In this case, the decision maker would have a myopic attitude and wants the cost to be minimized over a short period, whatever costs will be incurred later on. Whereas, if $\alpha \approx 0$, the decision maker considers the long-run behaviour of threat of the infrastructure. More accurately, we have the following theorem.

Definition 3.3.6. Let $\Gamma_\alpha = (S, A, D, \Lambda, P, \mathscr{K}, \mathscr{C}, \alpha)$, $\alpha > 0$, be finite-state finite-action CMDPs. A policy $\pi \in \Pi$ is *Blackwell optimal* if there is some $\alpha^* > 0$ such that π is optimal for all $0 < \alpha < \alpha^*$.

Theorem 3.3.7. *Let $\Gamma_\alpha = (S, A, D, \Lambda, P, \mathscr{K}, \mathscr{C}, \alpha)$, $\alpha > 0$, be finite-state finite-action CMDPs. Then there is a Blackwell optimal decision rule μ^*. Furthermore, μ^* is optimal for the expected average cost criterion.*

Proof. See (Puterman, 2005), Theorems 10.1.4 and 10.1.6. □

Example 3.3.8. In the airport example, we assume that there is a half life of 1 day or 24 hours respectively for the cost. That means, costs of the respective states and executed actions only count in half of their present amount 24 hours later. Here

$$e^{-24\alpha} = 0.5 \quad \Leftrightarrow \quad \alpha = \frac{\log(2)}{24} = 0.02888\ldots.$$

One could also define α to be the actual discount rate if the costs are monetary values.

3.3.7. The CMDP-Model

We assume that all exponentially distributed times for the occurrences of threat events and for the completions of elementary actions are independent for all sectors. In this case, we are able to formulate the CMDP from the parametrization given in sections 3.3.1–3.3.6. As already mentioned, the state space is

$$S = G^\Sigma,$$

and the action space is

$$A = A_0^\Sigma.$$

Let the number of the available security staff be $r \in \mathbb{N}$. We assume that each person of the security staff is able to perform exactly one elementary action in one sector at the same time. So, the decision maker is allowed to perform elementary actions different from the passive elementary action 0 in at most r sectors. Therefore, we define the restriction set

$$D(s) \coloneqq \left\{ a \in A \;\middle|\; \sum_{\sigma \in \Sigma} \left(1 - \delta_{0a(\sigma)}\right) \leq r \right\}, \quad s \in S,$$

where δ is the Kronecker delta. The transition rates of the CMDP are given by

$$\lambda(s,a) = \sum_{\sigma \in \Sigma} \left[\sum_{e \in \mathscr{E}(\sigma)} \lambda_e(s(\sigma)) + \lambda_\sigma(a(\sigma)) \right], \quad (s,a) \in D.$$

If $\lambda(s,a) > 0$, then the transition probabilities are explicitly given by

$$p^a_{ss'} = \frac{1}{\lambda(s,a)} \sum_{\sigma\in\Sigma} \left[\sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma)) \cdot \mathbb{1}\left(s' = \zeta_e(s)\right) + \lambda_\sigma(a(\sigma))\, \Phi^{s(\sigma)}_{a(\sigma),\sigma}(s'(\sigma)) \cdot \mathbb{1}\left(s' = \zeta^{s'(\sigma)}_{a(\sigma),\sigma}(s)\right) \right],$$

$s, s' \in S$, $a \in D(s)$, where $\mathbb{1}(\mathscr{S}) = 1$ if the mathematical statement $\mathscr{S}$ is true and $\mathbb{1}(\mathscr{S}) = 0$ if $\mathscr{S}$ is wrong due to the independence of all threat events and elementary actions. If $\lambda(s,a) = 0$, then the system remains in s forever so that s is absorbing under a, and $p^a_{ss'}$ can be defined arbitrarily. In our model, the setup costs are zero, i. e., $K(s,a) = 0$ for all $(s,a) \in D$. One might introduce set-up costs for the elementary actions. In this case, continuing an elementary action after a decision time, would result in another set-up cost. So continuing an action, cannot be modelled since the respective elementary action would be started anew. If the set-up cost is zero for all elementary actions, then setting up an elementary action again can be interpreted as continuing the elementary action due to the lack-of-memory property of the exponential distribution.

Now, we define the cost rates. Let $e \in \mathscr{E}(\sigma)$ for some $\sigma \in \Sigma$ and let $s \in S$. After an expected time of $1/\lambda_e(s(\sigma))$, e occurs, which yields the cost C_e. If we define the respective cost rate $\lambda_e(s(\sigma))C_e$, the cost C_e is incurred after the expected time $1/\lambda_e(s(\sigma))$ until occurrence of e. A similar argument for the cost rate is applied to the terms representing the removals of dangerous objects. Therefore, the cost rates of the CMDP are defined by

$$C(s,a) = \sum_{\sigma\in\Sigma} \left[\sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma))C_e + c_{a(\sigma)} + (1-\delta_{0a(\sigma)})\,\lambda_\sigma(a(\sigma))\,\gamma_\sigma(s(\sigma))C_\sigma \right], \quad (s,a) \in D,$$

since the cost rates of the elementary actions have to be added assuming that every elementary action except 0 is able to detect dangerous objects. The discount rate $\alpha > 0$ can be chosen arbitrarily.

To determine the contraction operator T for the expected total discounted cost criterion for the surveillance task with restriction set D, the transition probabilities P need not be calculated explicitly. For $v : S \to \mathbb{R}$, the operator T takes the form

$$\begin{aligned} Tv(s) = \min_{a\in D(s)} \Bigg\{ \frac{1}{\lambda(s,a)+\alpha} \Bigg[C(s,a) + \sum_{\sigma\in\Sigma} \Bigg[\sum_{e\in\mathscr{E}(s(\sigma))} \lambda_e(s(\sigma))\, v(\zeta_e(s)) \\ + \lambda_\sigma(a(\sigma)) \sum_{g'\in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}\, v\left(\zeta^{g'}_{a(\sigma),\sigma}(s)\right) \Bigg] \Bigg] \Bigg\}, \quad s \in S. \end{aligned} \tag{3.1}$$

The surveillance task consists in finding an optimal policy for the expected total discounted cost criterion. In other words, the surveillance task consists in finding a policy which minimizes the expected occurrences of (appropriately discounted) expensive threat events. Theoretically, as seen in section 2.3.2, this problem can be solved by linear programming which yields an optimal decision rule.

Next, we compute the number of states and actions depending on the number of sectors $|\Sigma|$ and the security staff size r in order to get some insight into the numerical complexity of the surveillance task.

Proposition 3.3.9. *For some infrastructure, let the parametrization be given as in sections 3.3.1–3.3.6 and let $r \in \mathbb{N}_0$. Then we have*

$$|S| = (g_{\max}+1)^{|\Sigma|} \quad \textit{and} \quad |A| = \sum_{R=0}^{r} \binom{|\Sigma|}{R} (|A_0|-1)^R.$$

Proof. The formula for the number of states is clear. If exactly $R \in \mathbb{N}_0$ elementary actions apart from 0 have to be executed, then there are $\binom{|\Sigma|}{R}$ subsets containing R sectors. In each sector, $|A_0|-1$ elementary actions apart from 0 can be executed making $(|A_0|-1)^R$ such actions for a given subset of R sectors. Adding over all $R = 0, \ldots, r$ yields the assertion. □

Example 3.3.10. With the parameters given in the accompanying example of this section where the security staff consists of two persons, we obtain the following optimal values:

$$v^*(0,0,0,0) = 505{,}180.3, \quad v^*(2,2,2,2) = 576{,}118.5, \quad v^*(4,4,4,4) = 2{,}766{,}976.2.$$

An optimal decision rule is depicted in Figure 3.2. It has to be read in the following way. The horizontal axis as well as the vertical axis illustrate the threat levels of two sectors. Take the horizontal axis. The two considered sectors are the

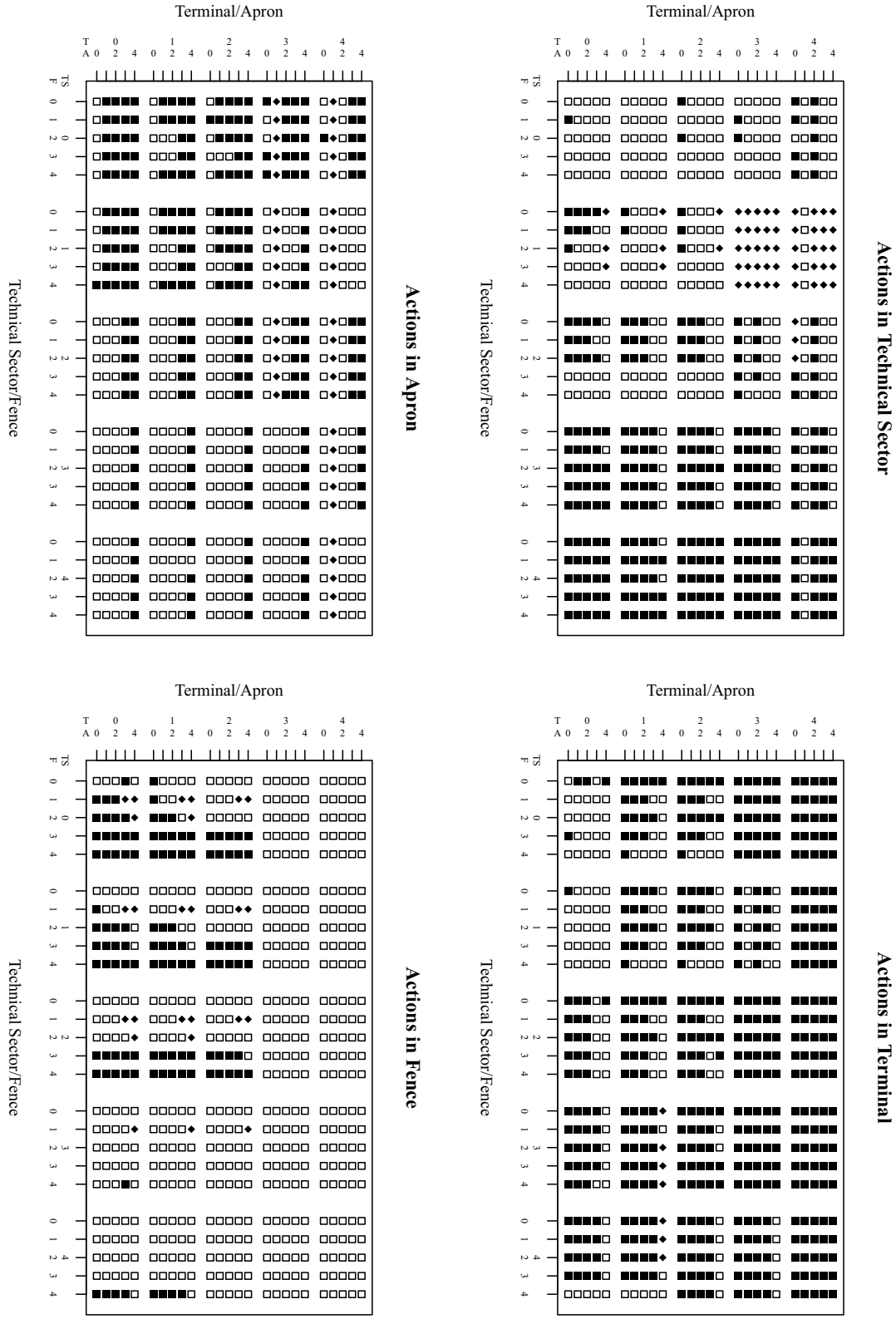


Figure 3.2.: Optimal decision rule for the accompanying example when the staff size is two.

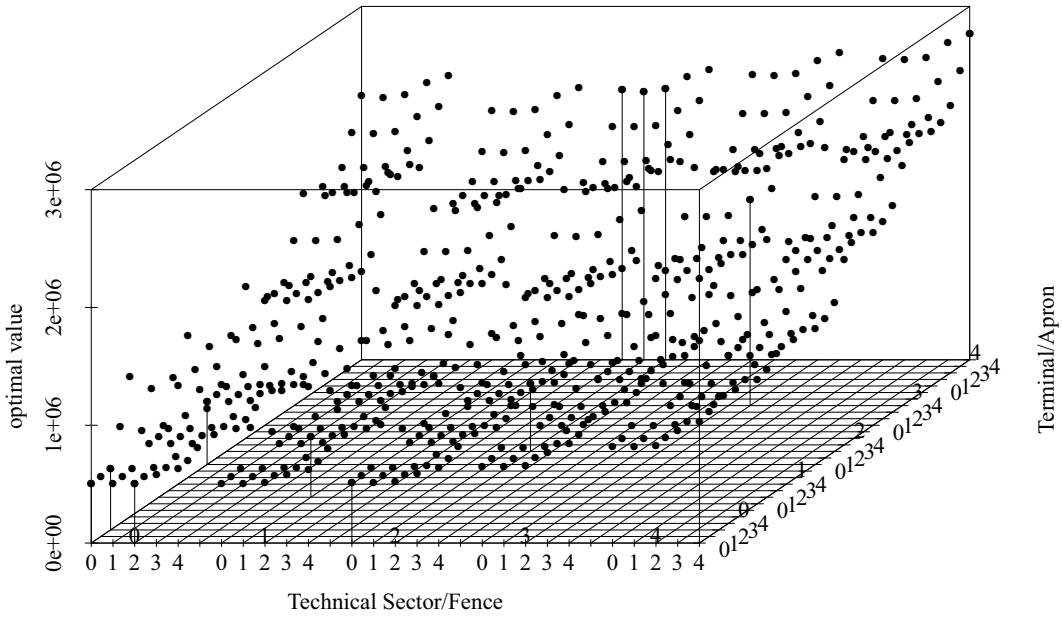


Figure 3.3.: Optimal value function of the accompanying example when the staff size is two.

technical sector and the fence. The first block of length five represents threat level 0 for the technical sector. Within this block, the threat level of the fence goes from 0 on the left side to 4 on the right side. The next block contains the states in which the threat level of the technical sector is 1. Again, within this block, the left column belongs to threat level 0 of the fence and the right column belongs to threat level 4 of the fence. In a similar manner, the vertical axis has to be interpreted. The blocks belong to the threat level of the terminal. The rows within such a block belong to the threat levels of the apron. Consider the top left mark. It belongs to the state s with $s(\mathrm{TS}) = 0$, $s(\mathrm{T}) = 4$, $s(\mathrm{A}) = 4$ and $s(\mathrm{F}) = 0$. The four large blocks of Figure 3.2 tell us which elementary action is optimal for the respective sector in the respective state. A white square corresponds to "Do nothing" (0), a black square standing on a vertex corresponds to "Camera evaluation" (1), and a black square standing on an edge corresponds to "Inspection walk" (2). In s, the optimal action $\mu^*(s)$ is given by $\mu^*(s)(\mathrm{TS}) = 0$, $\mu^*(s)(\mathrm{T}) = 2$, $\mu^*(s)(\mathrm{A}) = 2$ and $\mu^*(s)(\mathrm{F}) = 0$. This example shows that there is little structure in the optimal decision rule. Roughly, one could say that elementary actions besides 0 are optimal in expensive sectors or if the sector has many dependencies. But due to the lack of structure of an optimal decision rule, the decision maker cannot easily guess optimal actions.

The respective optimal value function is shown in Figure 3.3. The states are arranged in the x-y plane in the same manner as in the plot of the optimal decision rule. The states $(0,0,0,0)$, $(0,0,0,2)$, $(0,0,2,0)$, $(0,2,0,0)$, $(1,1,1,1)$, $(2,0,0,0)$, $(2,2,2,2)$, $(2,4,4,0)$, $(2,4,4,1)$, $(2,4,4,2)$, $(3,3,3,3)$ and $(4,4,4,4)$ are marked, where the order of the components is TS, T, A, F. From this illustration, it can be seen that the threat level of the terminal has the greatest influence on the optimal value function. The blocks corresponding to the same threat level of the terminal are more or less at the same height so that there are levels of the optimal value function. These levels are varied by the threat level of the apron. Whereas the influence of the threat levels of the technical sector and the fence is quite small. Note that, if we fix three threat levels of three sectors, then the optimal value function need not be increasing in the component of the last sector. This can be seen from the states $(2,4,4,s(\mathrm{F}))$, $s(\mathrm{F}) \in G$. We have $v^*(2,4,4,0) > v^*(2,4,4,1) < v^*(2,4,4,2)$.

Now, consider the same infrastructure where the staff size is four. The optimal values of the states $(0,0,0,0)$, $(2,2,2,2)$ and $(4,4,4,4)$ are

$$v^*(0,0,0,0) = 452{,}601.5, \quad v^*(2,2,2,2) = 493{,}369.6, \quad v^*(4,4,4,4) = 1{,}909{,}995.9.$$

The costs are smaller than in the case when the staff size is two. The reason is that more actions can be executed in if the staff size is four than when the staff size is two as shown in Theorem 4.1.1. An optimal decision rule is depicted in

Figure 3.4. Again, the optimal decision rule is low in structure. The optimal value function is illustrated in Figure 3.5. The level structure of the optimal value function is quite similar to the optimal function of the surveillance task with staff size two.

Also note that it need not be optimal to execute an elementary action when the current threat level is greater than 0. This is due to the rather non-transparent structure of the respective CMDP. Consider the state $(3,0,0,1)$ in which action $(2,2,0,0)$ is optimal. Why is it not optimal to execute 2 in F? We have $v^*(3,0,0,0) > v^*(3,0,0,1)$. Since executing 2 in F shifts some probability mass towards the state $(3,0,0,0)$, 2 is not optimal in F. A reason might be the modelling of elementary action 1. When a camera evaluation is completed in threat level 1, then there is quite a high probability to end up in threat level 0 so that the threat level of dependent sectors decreases by one with high probability. This mechanism could not be used if the threat level of the fence was 0.

3.4. Aspects of a Practical Application of the Model

In this section, we consider some aspects when modelling the dynamics of threat of an infrastructure in practice along the above lines.

In practice, various kinds of threat might appear in a sector of the underlying infrastructure. For example, in some sector, fire could break out or electric power could break down. But threat events concerning different threat types might have very different effects on the threat levels of dependent sectors. Furthermore, only the amount of threat with respect to a certain threat type might change so that the threat level which measures the occurrence rates of all threat events of all threat types cannot include this information. Potential fire outbreaks might threaten only adjoining sectors, whereas a potential breakdown of electric power might have effects on much more sectors. The model covers these thoughts by mapping one original sector of the infrastructure onto various sectors of the model where each sector stands for one threat type of the original sector only. Then the dynamics can exclusively be defined for each threat type. In doing so, the CMDP becomes more complex since the state space and the action space increase considerably.

If the set of threat events contains multiple threat types, then threat appears in a more or less abstract sense. So, the dynamics of threat has to be understood in a qualitative manner. One could think of the model as of a cloud of abstract threat spreading on the sectors of the infrastructure. The thicker the cloud gets, the more it costs. The cloud can be forced back by threat-reducing actions. But threat itself is not concrete so that no threat-specific countermeasures can be modelled.

It is not very hard to define meaningful parameters for most of the components of the threat model. A lot of parameters are known from a rigorous risk assessment of the infrastructure. For example, the structure of the infrastructure, threat events, the respective costs and possible elementary actions should be known. More tricky is the definition of the threat levels, where the occurrence rates of the threat events should be defined appropriately. The same applies to the transition functions and the transition mechanisms of the threat events and of the elementary actions. Both definitions are a matter of sure instinct. The discount factor can be chosen arbitrarily from a computational point of view, except for round-off errors. So, it should be chosen very small so that the long-run behaviour is considered in view of Theorem 3.3.7.

In order to provide support for the correct risk assessment of the security staff, the current threat state might be illustrated on a monitor by depicting a risk map with several colours indicating the various threat levels of the sectors. Furthermore, optimal actions might be shown on the map in order to provide decision support to the security staff.

The main problem of the model is its complexity. For an infrastructure consisting of five sectors where $g_{\max} = 4$, the exact optimal decision rule is computable by linear programming. But for a similar infrastructure consisting of six sectors, the solution of the respective linear program was not available within 48 hours. This is a very small infrastructure compared to realistic infrastructures like a small airport. To tackle the surveillance task, we have to consider approximation techniques. This is done in chapter 5.

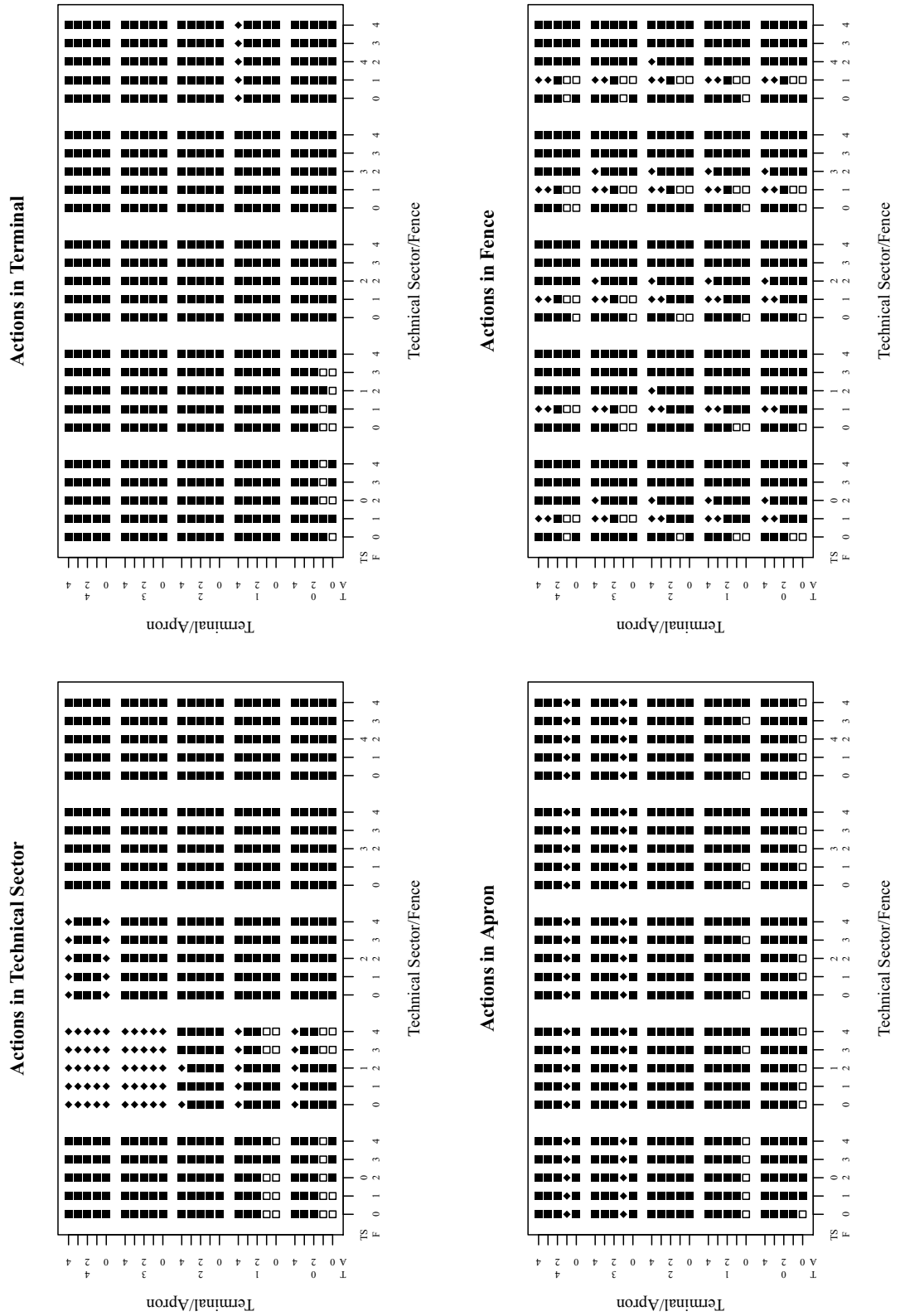


Figure 3.4.: Optimal decision rule for the accompanying example when the staff size is four.

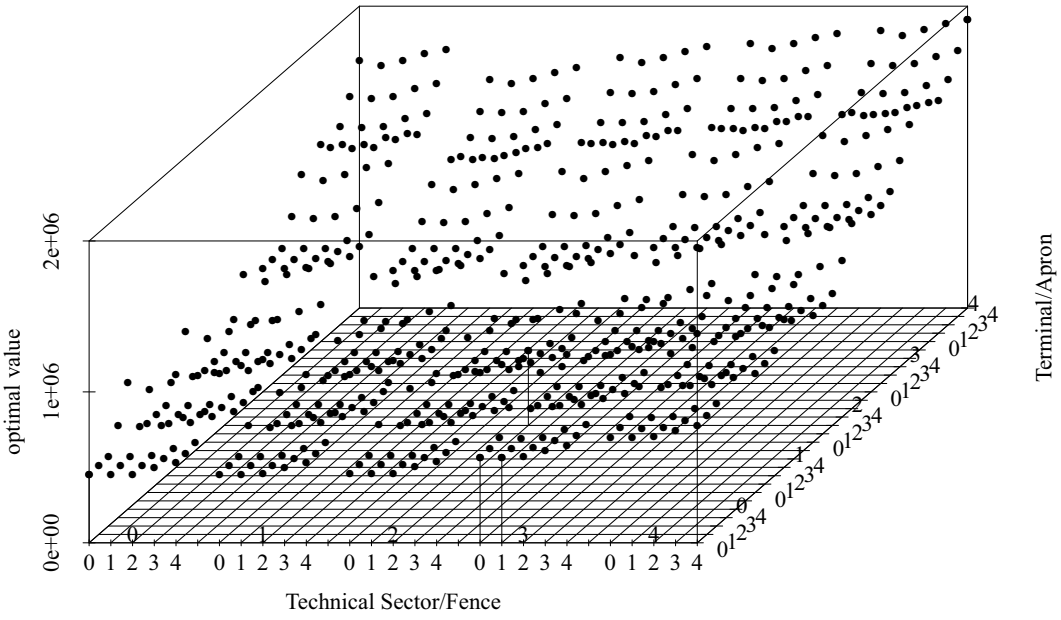


Figure 3.5.: Optimal value function of the accompanying example whe the staff size is four.

4. Model Analysis

In this chapter, we consider several aspects of the mathematical model for the surveillance task. At first, we show that increasing the staff size leads to a decrease of the optimal value. Then we derive that a control threshold policy is optimal in the case of an infrastructure consisting of exactly one sector if it satisfies certain assumptions. This is followed by a section which considers infrastructures without dependencies. In the second section, we consider how a policy derived from an infrastructure with no dependencies performs in the very same infrastructure, now, with dependencies. This is done by means of a generalization of the threat model, which we call a coupled CMDP.

4.1. Exact Solution

In this section, we derive some results concerning the exact solution of the surveillance task. At first, we show that it pays off to have a maximal number of security staff at hand. In a second part, we demonstrate that a control threshold policy is optimal for a surveillance task for a one-sector infrastructure. Third, it is illustrated that it is optimal to solve the surveillance task for each sector separately if the infrastructure has no dependencies.

4.1.1. Dependency of the Optimal Value Function on the Staff Size

At first, we show that increasing the staff size leads to a decrease of the optimal value so that it is profitable to have maximal staff size at hand.

Theorem 4.1.1. *For some infrastructure, let the parametrization be given as in chapter 3. Let $r \leq r'$, and let v_r^* and $v_{r'}^*$ be the optimal value functions for the surveillance task with staff size r and r' respectively. Then*

$$v_r^* \geq v_{r'}^*.$$

Proof. Let D_r and $D_{r'}$ denote the restriction sets for the surveillance task with staff size r and r' respectively. Moreover, let T_r and $T_{r'}$ be the respective contraction operators as defined in (3.1). Now, let $v : S \to \mathbb{R}$ be arbitrary. Since $D_r \subset D_{r'}$, we have $T_r v \geq T_{r'} v$, from which $T_r^2 v \geq T_r T_{r'} v \geq T_{r'}^2 v$ follows from isotony of T_r. By induction, it follows $T_r^n v \geq T_{r'}^n v$ for all $n \in \mathbb{N}$. Since $T_r^n v \to v_r^*$, $n \to \infty$, and $T_{r'}^n \to v_{r'}^*$, $n \to \infty$, by value iteration of Proposition 2.3.6, the assertion follows. □

So, it is best having a large security staff in readiness, as long as only the effective working time has to be paid. An example of this theorem has already been given in Example 3.3.10. In practice, there is a fixed number of personnel r employed so that, in fact, actions do not cost anything. But each employee earns a fixed cost rate $c_{\text{staff}} \geq 0$ per time unit. In this case, the arising discounted cost $r\, c_{\text{staff}}/\alpha$ for the salaries has to be added to the optimal value function v_r^* so that the real expected total discounted cost for $s \in S$ is $v_r^*(s) + r\, c_{\text{staff}}/\alpha$. As deduced in Theorem 4.1.1, the optimal value function v_r^* is decreasing for r increasing. But the additional amount $r\, c_{\text{staff}}/\alpha$ might overcompensate this benefit. In this manner, the threat model could be used as a tool for optimal resource planning by choosing an r^* such that $v_{r*}^*(s) + r^*\, c_{\text{staff}}/\alpha$ is minimal for a given state $s \in S$.

4.1.2. Optimality of a Control Threshold Policy

Although there is no practical interest, we examine the case of an infrastructure consisting of only one sector next. We derive the optimality of a control threshold policy (CTP) if the parametrization of the threat model satisfies certain assumptions. When there are two available actions, a CTP chooses the first action for every state below some control threshold and uses the second action for every state above this control threshold.

CTPs turn out to be optimal in several applications. For example, (Economou, 2003; Kyriakidis, 2006) derive CTPs for their respective CMDP-models. (Gupta and Wang, 2008) consider among others the revenue management of a single-physician clinic. Patients have different preferences whether they want a same-day appointment or a scheduled future appointment. Their model is a finite-horizon MDP, and they derive the optimality of a CTP, which is to let choose the patient an arbitrary booking slot up to some booking limit. If the booking limit is reached, then the appointment request should be declined.

Throughout this section, let $\Sigma = \{\sigma\}$ and $G = \{0, \ldots, g_{\max}\}$ so that the state space is $S = G$. In the following, we omit the index σ whenever possible. Then we have the non-empty finite set of threat events $\mathscr{E}$. We assume that the action

space is $A = A_0 = \{0,2\}$. In terms of the accompanying example of the previous chapter, this means that the decision maker is able to initiate the action "Do nothing" or the action "Inspection walk." The transition mechanism of action 2 is given by $\Phi_2^g(g')$, $g, g' \in G$, the cost is given by $c_2 \geq 0$, and the rate is given by $\lambda(2) > 0$. Since there is only one sector, we need not define parameters for dependent sectors.

As we shall see, the term

$$Hv(g) := \lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(g) \left[C_e + v(\Psi_e(g)) - \frac{c_2}{\lambda(2)} - \gamma(g) C_\sigma - \Phi_2 v(g) \right],$$

where $\Phi_2 v(g) := \sum_{g' \in G} \Phi_2^g(g')\, v(g')$, $g \in G$, and $v : S \to \mathbb{R}$, plays an important role in the following analysis. The value $\Phi_2 v(g)$ is the expectation of $v(X)$, where X is the random variable which is the result of action 2 when the current state is g. For now, let us make the following assumption.

Assumption 4.1.2. Let the parametrization of an infrastructure with $\Sigma = \{\sigma\}$ be given according to chapter 3. Moreover, let the parameters satisfy the following assumptions:

1. Let γ be increasing.
2. Let Ψ_e be increasing for all $e \in \mathscr{E}$.
3. Let λ_e be increasing for all $e \in \mathscr{E}$.
4. Let $\lambda_e(g)/\lambda(g,0)$ be increasing in g for every $e \in \mathscr{E}$.
5. Let $\Phi_2 v$ be increasing for all increasing $v : S \to \mathbb{R}_{\geq 0}$.
6. Let $v : G \to \mathbb{R}_{\geq 0}$ be increasing. If $Hv(g) > \alpha\,[c_2 + \lambda(2)\,\gamma(g)\,C_\sigma + \lambda(2)\,\Phi_2 v(g)]$ for some $g \in G$, then it holds $Hv(g') > \alpha\,[c_2 + \lambda(2)\,\gamma(g)\,C_\sigma + \lambda(2)\,\Phi_2 v(g')]$ for all $g' \geq g$.
7. Let $v : G \to \mathbb{R}_{\geq 0}$ be increasing.
   a) If $g > 0$ and $Hv(g) \geq 0$, then it holds $Hv(g') \leq Hv(g'')$ for all $g - 1 \leq g' \leq g''$.
   b) If $Hv(0) \geq 0$, then it holds $Hv(0) \leq Hv(g')$ for all $g' \geq 0$.

In practice, Assumptions 4.1.2.1–3 are plausible assumptions for an adequate threat model. It is more likely to find a dangerous object, and it is more likely that threat events occur more often in a sector which is in a high threat level than in a sector which is in a low threat level. Furthermore, when a threat event occurs in a sector with a high threat level, the subsequent threat level will be higher than when the threat event occurs in a sector with a low threat level. Assumption 4.1.2.4 is a technical assumption. From Assumption 4.1.2.5, it follows that the expected threat level after action 2 is accomplished is increasing in g. Assumptions 4.1.2.6 and 7 are implicit assumptions on action 2, on the threat events and on γ. Later, in Assumption 4.1.5, we state simple conditions under which Assumption 4.1.2 holds. We have the following theorem.

Theorem 4.1.3. *Under Assumption 4.1.2, the optimal value function v^* is increasing, and a CTP μ^* is optimal, i. e.,*

$$\mu^*(g) = \begin{cases} 0, & \text{if } g \leq g_{\text{threshold}} \\ 2, & \text{if } g > g_{\text{threshold}} \end{cases}, \quad g \in G,$$

for some $g_{\text{threshold}} \in G \cup \{-1\}$. (If $g_{\text{threshold}} = -1$, then action 2 is optimal in every state.)

Proof. At first, note that $\lambda(g,0) = \sum_{e \in \mathscr{E}} \lambda_e(g)$ and $\lambda(g,2) = \lambda(g,0) + \lambda(2)$, $g \in G$. We define the respective one-step operators of the actions 0 and 2 for $v : G \to \mathbb{R}$ by

$$T_0 v(g) := \frac{\sum_{e \in \mathscr{E}} \lambda_e(g) C_e}{\lambda(g,0) + \alpha} + \frac{\lambda(g,0)}{\lambda(g,0) + \alpha} \sum_{e \in \mathscr{E}} \frac{\lambda_e(g)}{\lambda(g,0)}\, v(\Psi_e(g)) = \frac{\sum_{e \in \mathscr{E}} \lambda_e(g)\,[C_e + v(\Psi_e(g))]}{\lambda(g,0) + \alpha},$$

$$\begin{aligned} T_2 v(g) &:= \frac{\sum_{e \in \mathscr{E}} \lambda_e(g) C_e + c_2 + \lambda(2)\,\gamma(g) C_\sigma}{\lambda(g,2) + \alpha} + \frac{\lambda(g,2)}{\lambda(g,2) + \alpha} \left[\sum_{e \in \mathscr{E}} \frac{\lambda_e(g)}{\lambda_e(g,2)}\, v(\Psi_e(g)) + \frac{\lambda(2)}{\lambda(g,2)} \sum_{g' \in G} \Phi_2^g(g')\, v(g') \right] \\ &= \frac{\sum_{e \in \mathscr{E}} \lambda_e(g)\,[C_e + v(\Psi_e(g))] + c_2 + \lambda(2)\,\gamma(g) C_\sigma + \lambda(2)\,\Phi_2 v(g)}{\lambda(g,2) + \alpha} \end{aligned}$$

respectively for $g \in G$. We define $Tv := \min\{T_0 v, T_2 v\}$ so that $v^* = Tv^*$ by Theorem 2.3.4. For $v : G \to \mathbb{R}$ and $g \in G$, we have

$$\begin{aligned} & T_0 v(g) > T_2 v(g) \\ \Leftrightarrow \quad & \lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(g) \left[C_e + v(\Psi_e(g))\right] > (\lambda(g,0) + \alpha) \left[c_2 + \lambda(2)\, \gamma(g) C_\sigma + \lambda(2)\, \Phi_2 v(g)\right] \qquad (4.1) \\ \Leftrightarrow \quad & Hv(g) > \alpha \left[c_2 + \lambda(2)\, \gamma(g) C_\sigma + \lambda(2)\, \Phi_2 v(g)\right]. \qquad (4.2) \end{aligned}$$

Let v be non-negative and increasing. Let (4.2) be valid for some $g \in G$. From Assumption 4.1.2.6, it follows

$$\alpha \left[c_2 + \lambda(2)\, \gamma(g') C_\sigma + \lambda(2)\, \Phi_2 v(g')\right] < Hv(g')$$

for all $g' \geq g$. Hence, $T_2 v(g') = Tv(g')$ for all $g' \geq g$. Next, we want to show that if $v : G \to \mathbb{R}_{\geq 0}$ is increasing, then Tv is increasing, too. This is done in three steps:

1. First, we show that $T_0 v$ is increasing. To this end, we define for $g \in G$ the random variable X_g with distribution

$$P(X_g = 0) = \frac{\alpha}{\lambda(g,0) + \alpha}, \quad P(X_g = C_e + v(\Psi_e(g))) = \frac{\lambda_e(g)}{\lambda(g,0) + \alpha}, \quad e \in \mathscr{E}.$$

   Then $EX_g = T_0 v(g)$. Furthermore, from Assumptions 4.1.2.3 and 4.1.2.4 we obtain that $\lambda_e(g)/(\lambda(g,0) + \alpha)$ is increasing in g for every $e \in \mathscr{E}$. (Let $e \in \mathscr{E}$ and $g' \geq g$, then $\lambda_e(g')/\lambda(g',0) \geq \lambda_e(g)/\lambda(0,g) \Leftrightarrow \lambda_e(g')\,\lambda(g,0) \geq \lambda(g)\,\lambda(g',0) \Rightarrow \lambda_e(g')\,(\lambda(g,0) + \alpha) \geq \lambda_e(g)\,(\lambda(g',0) + \alpha)$.) Therefore, if $g' \geq g$, then $X_{g'}$ is stochastically larger than X_g due to Assumption 4.1.2.2. Hence $T_0 v(g') = EX_{g'} \geq EX_g = T_0 v(g)$.

2. The next assertion we want to show is the following: if $T_0 v(g) > T_2 v(g)$ for some $g > 0$, then $T_2 v(g') \leq T_2 v(g'')$ for all $g - 1 \leq g' \leq g''$. Loosely speaking, we want to show: if 2 is better than 0 for some $g > 0$, then $T_2 v$ increases already at $g - 1$. To this end, let $g - 1 \leq g' < g''$. At first, since $T_0 v(g) > T_2 v(g)$, we have $Hv(g) > \alpha \left[c_2 + \lambda(2)\, \gamma(g) C_\sigma + \lambda(2)\, \Phi_2 v(g)\right] \geq 0$ by (4.2) since all parameters are non-negative. By Assumption 4.1.2.7a, we have $Hv(g'') \geq Hv(g')$, which holds if and only if

$$\begin{aligned} & \lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(g'') \left[C_e + v(\Psi_e(g'')) - \frac{c_2}{\lambda(2)} - \gamma(g'') C_\sigma - \Phi_2 v(g'')\right] \\ & \quad - \lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(g') \left[C_e + v(\Psi_e(g')) - \frac{c_2}{\lambda(2)} - \gamma(g') C_\sigma - \Phi_2 v(g')\right] \geq 0 \\ \Leftrightarrow \quad & \lambda(2) \sum_{e \in \mathscr{E}} \left[\lambda_e(g'') \left[C_e + v(\Psi_e(g'')) - \frac{c_2}{\lambda(2)} - \gamma(g'') C_\sigma - \Phi_2 v(g'')\right]\right. \\ & \qquad \left. - \lambda_e(g') \left[C_e + v(\Psi_e(g')) - \frac{c_2}{\lambda(2)} - \gamma(g') C_\sigma - \Phi_2 v(g')\right]\right] \geq 0. \end{aligned} \qquad (4.3)$$

   Furthermore, we have $T_2 v(g') \leq T_2(g'')$ if and only if

$$\begin{aligned} & (\lambda(g'',0) + \lambda(2) + \alpha) \left[c_2 + \lambda(2)\, \gamma(g') C_\sigma + \lambda(2)\, \Phi_2 v(g') + \sum_{e \in \mathscr{E}} \lambda_e(g') \left[C_e + v(\Psi_e(g'))\right]\right] \\ & \quad \leq (\lambda(g',0) + \lambda(2) + \alpha) \left[c_2 + \lambda(2)\, \gamma(g'') C_\sigma + \lambda(2)\, \Phi_2 v(g'') + \sum_{e \in \mathscr{E}} \lambda_e(g'') \left[C_e + v(\Psi_e(g''))\right]\right] \\ \Leftrightarrow \quad & \lambda(g',0) \left[c_2 + \lambda(2)\, \gamma(g'') C_\sigma + \lambda(2)\, \Phi_2 v(g'') + \sum_{e \in \mathscr{E}} \lambda_e(g'') \left[C_e + v(\Psi_e(g''))\right]\right] \\ & \quad - \lambda(g'',0) \left[c_2 + \lambda(2)\, \gamma(g') C_\sigma + \lambda(2)\, \Phi_2 v(g') + \sum_{e \in \mathscr{E}} \lambda_e(g') \left[C_e + v(\Psi_e(g'))\right]\right] \\ & \quad + \lambda(2) \sum_{e \in \mathscr{E}} \left[\lambda_e(g'') \left[C_e + v(\Psi_e(g''))\right] - \lambda_e(g') \left[C_e + v(\Psi_e(g'))\right]\right] \\ & \quad + \alpha \sum_{e \in \mathscr{E}} \left[\lambda_e(g'') \left[C_e + v(\Psi_e(g''))\right] - \lambda_e(g') \left[C_e + v(\Psi_e(g'))\right]\right] \end{aligned}$$

$$
\begin{aligned}
&+(\lambda(2)+\alpha)\,\lambda(2)\,\left[C_\sigma\left(\gamma(g'')-\gamma(g')\right)+\Phi_2 v(g'')-\Phi_2 v(g')\right]\geq 0\\
\Leftrightarrow\quad &\underbrace{\lambda(g',0)\sum_{e\in\mathscr{E}}\lambda_e(g'')\left[C_e+v(\Psi_e(g''))\right]-\lambda(g'',0)\sum_{e\in\mathscr{E}}\lambda_e(g')\left[C_e+v(\Psi_e(g'))\right]}_{\geq 0,\text{ since }\lambda_e(g)/\lambda(g,0),\,\Psi_e(g),\,v(g)\text{ are increasing by Assumptions 4.1.2.2 and 4}}\\
&+\lambda(2)\sum_{e\in\mathscr{E}}\Bigg[\lambda_e(g'')\left[C_e+v(\Psi_e(g''))-\frac{c_2}{\lambda(2)}-\gamma(g'')C_\sigma-\Phi_2 v(g)\right]\\
&\qquad\underbrace{-\lambda_e(g')\left[C_e+v(\Psi_e(g'))-\frac{c_2}{\lambda(2)}-\gamma(g')C_\sigma-\Phi_2 v(g)\right]\Bigg]}_{\geq 0,\text{ due to (4.3)}}\\
&\underbrace{+\alpha\sum_{e\in\mathscr{E}}\left[\lambda_e(g'')\left[C_e+v(\Psi_e(g''))\right]-\lambda_e(g')\left[C_e+v(\Psi_e(g'))\right]\right]}_{\geq 0,\text{ since }\lambda_e(g),\,\Psi_e(g),\,v(g)\text{ are increasing by Assumptions 4.1.2.3 and 2}}\\
&\underbrace{+(\lambda(2)+\alpha)\,\lambda(2)\,\left[C_\sigma\left(\gamma(g'')-\gamma(g')\right)+\Phi_2 v(g'')-\Phi_2 v(g')\right]}_{\geq 0,\text{ since }\gamma\text{ and }\Phi_2 v\text{ are increasing by Assumptions 4.1.2.1 and 5}}\geq 0.
\end{aligned}
$$

So, the last statement is true, and so is the assertion.

3. Now, we show a similar assertion as assertion 2 for $g=0$: if $T_0v(0)>T_2v(0)$, then $T_2v(g')\leq T_2v(g'')$ for all $g'\leq g''$. The proof of this assertion is similar to the proof of assertion 2, the only difference being that Assumption 4.1.2.7b is used and that only $g'\geq 0$ are considered.

From 1–3, we conclude that $Tv=\min\{T_0v,T_2v\}$ is increasing. Furthermore $Tv\geq 0$. Iterating steps 1–3, implies that v, Tv, T^2v, …, are increasing and non-negative. By value iteration of Proposition 2.3.6, we obtain that $v^*=\lim_{n\to\infty}T^nv$ is increasing and non-negative. Plugging in v^* for v in (4.2), we see that the right-hand side is non-negative. Assumption 4.1.2.7 implies: if action 2 is optimal in g, then 2 is optimal for all $g'\geq g$. Thus, a CTP is optimal. □

Knowing that a CTP is optimal for a certain problem, reduces the complexity of computation of an optimal decision rule since the space of decision rules that has to be considered is drastically reduced. In our case, instead of considering $2^{g_{\max}+1}$ decision rules, it remain $g_{\max}+2$ CTPs. Unfortunately, there is no improvement for the linear programming method since all constraints concerning the restriction set have to be used. One possibility to exploit the CTP-structure of an optimal policy is to search all CTPs and pick one with minimal expected total discounted cost. Howard's policy improvement algorithm could also be speeded up according to Algorithm 4.1: in the algorithm, the policy $\mu_k(g^k_{\text{threshold}})$ is defined by $\mu_k(g^k_{\text{threshold}})(g)=0$, $g\leq g^k_{\text{threshold}}$, and $\mu_k(g^k_{\text{threshold}})(g)=2$, $g>g^k_{\text{threshold}}$, for some control threshold $g^k_{\text{threshold}}\in G$, $k=0,1,\ldots$.

Proposition 4.1.4. *Algorithm 4.1 terminates after a finite number of steps K and μ_K is optimal.*

Proof. By the proof of Theorem 4.1.3, we have that $T_{\mu(g)}v(g)=Tv(g)$ for all $g\in G$ for some CTP μ for every increasing $v:G\to\mathbb{R}_{\geq 0}$. For $k=0,1,\ldots,K$, condition (4.4) holds if and only if $T_2v^{\mu_k}(g)<T_0v^{\mu_k}(g)$ by (4.1). Hence, a policy improvement step is performed in the sense of Howard's policy improvement algorithm. The same is true if v^{μ_k} is decreasing. So, Algorithm 4.1 is equivalent to Howard's improvement algorithm (cf. (Puterman, 2005), Theorem 6.4.2), from which the assertion follows. □

Due to line 6 of Algorithm 4.1, one need not consider all states of the infrastructure in the improvement step if v^{μ_k} is increasing which increases the speed of Howard's improvement algorithm in general. Assumption 4.1.2 holds under the following conditions:

Assumption 4.1.5. Let the parametrization of an infrastructure with $\Sigma=\{\sigma\}$ be given according to chapter 3. Moreover, let the parameters satisfy the following assumptions:

1. Let $\gamma\equiv 0$.
2. Let Ψ_e be increasing for all $e\in\mathscr{E}$.
3. Let $\lambda_e(g)=c(g)\,\lambda_e(0)$ for all $g\in G$ and $e\in\mathscr{E}$ where $c(g)\geq 1$ is increasing in g.
4. Let $\Phi_2^g(0)=1$ for all $g\in G$.

Algorithm 4.1 Policy improvement algorithm.

Require: parametrization of the infrastructure according to chapter 3 and Assumption 4.1.2

define $k = 0$ and $v^{\mu_{-1}} = -\infty$
choose $g^0_{\text{threshold}} \in G$ and determine $\mu_0(g^0_{\text{threshold}})$
determine v^{μ_0}
while $v^{\mu_k} \neq v^{\mu_{k-1}}$ **do**
  if v^{μ_k} is increasing **then**
    determine smallest $g \in G$ such that

$$\lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(g) \left[C_e + v^{\mu_k}(\Psi_e(g))\right] - (\lambda(g,0) + \alpha)\left(c_2 + \lambda(2)\,\gamma(\sigma)\,C_\sigma + \lambda(2)\,\Phi_2 v^{\mu_k}(g)\right) > 0 \tag{4.4}$$

    if such a g exists, define $g^{k+1}_{\text{threshold}} = g - 1$ and $\mu_{k+1}(g^{k+1}_{\text{threshold}})$
    if no such g exists **then**
      define $g^{k+1}_{\text{threshold}} = g_{\max}$ and $\mu_{k+1}(g^{k+1}_{\text{threshold}})$
    end if
  else
    determine μ_{k+1} such that $\mu_{k+1}(g) \in \arg\min_{a \in \{0,2\}} \{T_a v^{\mu_k}(g)\}$ for all $g \in G$
  end if
  $k \leftarrow k + 1$
  determine v^{μ_k}
end while
return μ_k is optimal

Assumption 4.1.5.1 means that there are no dangerous objects or that removing of dangerous objects is free of cost respectively. Assumption 4.1.5.2 again says that threat events occurring at higher threat levels leave the sector at a higher threat level than when occurring at a lower threat level. Assumption 4.1.5.3 has the interpretation that occurrences of threat events increase equally for every threat level. By Assumption 4.1.5.4, the inspection walk is carried out perfectly and therefore leaving the sector at threat level 0 when it is accomplished.

Corollary 4.1.6. *Under Assumption 4.1.5, the optimal value function v^* is increasing and a CTP is optimal. Furthermore, action* 0 *is optimal in state* 0*.*

Proof. We have to verify Assumptions 4.1.2.1–7. Assumptions 4.1.2.1–3 are trivially satisfied. Since

$$\frac{\lambda_e(g)}{\lambda(g,0)} = \frac{\lambda_e(g)}{\sum_{\tilde{e} \in \mathscr{E}} \lambda_{\tilde{e}}(g)} = \frac{c(g)\,\lambda_e(0)}{c(g) \sum_{\tilde{e} \in \mathscr{E}} \lambda_{\tilde{e}}(0)} = \frac{\lambda_e(0)}{\lambda(0,0)}$$

is constant for all $g \in G$ and $e \in \mathscr{E}$, Assumption 4.1.2.4 is satisfied. Assumption 4.1.2.5 holds since $\Phi_2 v(g) = v(0)$, $g \in G$, is constant. Assumption 4.1.2.6 is satisfied since

$$\alpha\left[c_2 + \lambda(2)\,\gamma(g)\,C_\sigma + \lambda(2)\,\Phi_2 v(g)\right] = \alpha\left[c_2 + \lambda(2)\,v(0)\right]$$

is constant for all $g \in G$ as soon as Assumptions 4.1.2.7a and 7b are shown, which follows. Let $v : G \to \mathbb{R}_{\geq 0}$ be increasing $Hv(g) \geq 0$ for some $g \in G$. Then $\sum_{e \in \mathscr{E}} \lambda_e(0)\left[C_e + v(\Psi_e(g)) - c_2/\lambda(2) - v(0)\right] \geq 0$, which is increasing in g. For $g' \geq g$, we then have

$$\begin{aligned} 0 \leq Hv(g) &= c(g)\,\lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(0) \left[C_e + v(\Psi_e(g)) - \frac{c_2}{\lambda(2)} - v(0)\right] \\ &\leq c(g)\,\lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(0) \left[C_e + v(\Psi_e(g')) - \frac{c_2}{\lambda(2)} - v(0)\right] \leq c(g')\,\lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(0) \left[C_e + v(\Psi_e(g')) - \frac{c_2}{\lambda(2)} - v(0)\right] \\ &= Hv(g'). \end{aligned} \tag{4.5}$$

If $g > 0$, we also have

$$Hv(g) = c(g)\,\lambda(2) \sum_{e \in \mathscr{E}} \lambda_e(0) \left[C_e + v(\Psi_e(g)) - \frac{c_2}{\lambda(2)} - v(0)\right]$$

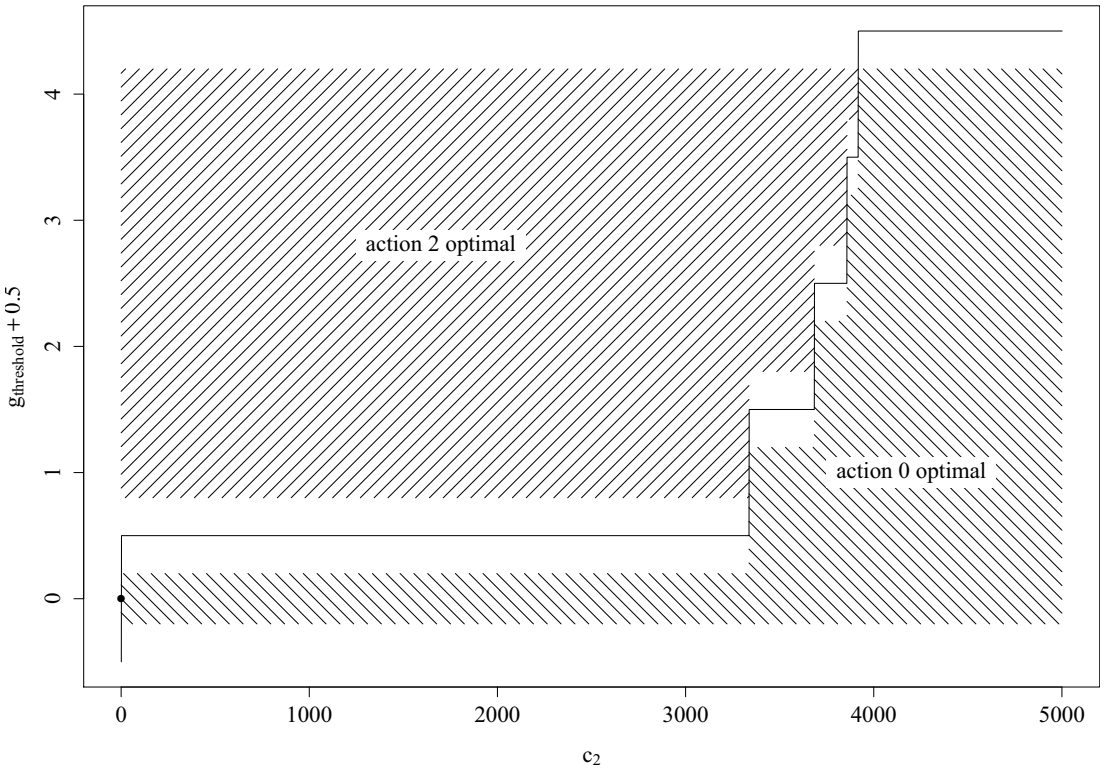


Figure 4.1.: Control threshold dependent on the cost rate c_2.

$$\geq c(g-1)\,\lambda(2)\sum_{e\in\mathscr{E}}\lambda_e(0)\left[C_e+v(\Psi_e(g-1))-\frac{c_2}{\lambda(2)}-v(0)\right]=Hv(g-1) \tag{4.6}$$

since $c(g)$, $c(g-1)$, $\lambda(2)\geq 0$ and $\sum_{e\in\mathscr{E}}\lambda_e(0)\,[C_e+v(\Psi_e(g))-c_2/\lambda(2)-v(0)]\geq 0$ so that "$\geq$" in (4.6) holds independent of the sign of $\sum_{e\in\mathscr{E}}\lambda_e(0)\,[C_e+v(\Psi_e(g-1))-c_2/\lambda(2)-v(0)]$. Hence, $Hv(g)\geq Hv(g-1)$. So, the first assertion follows from Theorem 4.1.3. Under Assumption 4.1.5, optimality of action 0 in state 0 follows in the following way: by the Bellman equation, action 2 is optimal in state 0 if and only if $T_2v^*(0)=v^*(0)$, which holds if and only if

$$\sum_{e\in\mathscr{E}}\lambda_e(0)\,[C_e+v^*\,(\Psi_e(0))]+c_2+\lambda(2)\,v^*(0)=(\lambda(0,0)+\lambda(2)+\alpha)v^*(0)\quad\Leftrightarrow\quad T_0v^*(0)+c_2=v^*(0) \tag{4.7}$$

by the definitions of T_0 and T_2 in the proof of Theorem 4.1.3. So if $c_2=0$, then actions 0 and 2 both are optimal. Let $c_2>0$ and assume that action 2 is optimal in state 0. Then it follows

$$T_0v^*(0)\geq v^*(0)\overset{(4.7)}{=}T_0v^*(0)+c_2>T_0v^*(0),$$

which is a contradiction. Hence, action 0 is optimal in state 0. □

Remark 4.1.7. Note that if $c_2=0$, then under Assumption 4.1.5, action 0 and action 2 both are optimal in state 0.

Example 4.1.8. We consider a one-sector infrastructure $\{\sigma\}$ satisfying Assumption 4.1.5. The parameters can be looked up in the appendix in section B.1. In this example, we consider several cost rates for action 2, $c_2\in\{0,1,\ldots,5000\}$, and compute the control threshold $g_{\text{threshold}}$ for the respective surveillance tasks. The result is presented in Figure 4.1. The solid line indicates $g_{\text{threshold}}+0.5$ depending on c_2. In states which are below this line, action 0 is optimal, whereas in states above this line, action 2 is optimal. The control threshold $g_{\text{threshold}}$ is increasing with c_2. The single dot in $(0,0)$ represents the optimality of action 2 in threat level 0 if $c_2=0$. For $c_2=0$, we have $g_{\text{threshold}}=-1$, which is not a contradiction to Corollary 4.1.6 since action 0 and action 2 both are optimal in 0.

Remark 4.1.9. Assumptions under which a CTP for general MDPs is optimal are derived, e. g., in (Altman and Stidham, Jr., 1995). Their proof of optimality of a CTP requires tedious calculations where the notion of submodularity is crucial for their analysis. It is done by considering finite-horizon problems and deriving optimal policies for these problems. By letting the time horizon to infinity, submodularity is derived for the infinite-horizon model, from which the optimality of a CTP is concluded.

4.1.3. Optimal Policies for Infrastructures without Dependencies and without Resource Restrictions

In this section, we consider an infrastructure consisting of several sectors which has no dependencies, i. e., $N \equiv 0$. Furthermore, we assume that the restriction set satisfies $D = A$ so that there are no restrictions on the actions, meaning that the security staff size is $|\Sigma|$. So, the infrastructure can be seen as various independent projects, where each sector represents a single project.

Definition 4.1.10. Let $(\mathrm{CMDP}_i) = (S_i, A_i, D_i, \Lambda_i, P_i, \mathscr{K}_i, \mathscr{C}_i, \alpha_i)$, $i = 1, \ldots, n$, be CMDPs as defined in Definition 2.1.1 such that $\mathscr{K}_i \equiv 0$, $i = 1, \ldots, n$, and such that they have the same discount rate $\alpha_i = \alpha > 0$. Then we define the *compound CMDP* $(\mathrm{CMDP}_i)_{i=1,\ldots,n} = (S, A, D, \Lambda, P, \mathscr{K}, \mathscr{C}, \alpha)$ by

$$S = \bigtimes_{i=1}^{n} S_i, \quad A = \bigtimes_{i=1}^{n} A_i, \quad D(s_1, \ldots, s_n) = \bigtimes_{i=1}^{n} D_i(s_i), \quad K(s,a) = 0, \quad C(s,a) = \sum_{i=1}^{n} C_i(s_i, a_i),$$

$$\lambda(s,a) = \sum_{i=1}^{n} \lambda_i(s_i, a_i), \quad p^{(a_1,\ldots,a_n)}_{(s_1,\ldots,s_n)(s_1,\ldots,s_{j-1},s'_j,s_{j+1},\ldots,s_n)} = \frac{\lambda_j(s_j, a_j)}{\lambda(s,a)} p^{a_j}_{j,s_j s'_j},$$

$s = (s_1, \ldots, s_n) \in S$, $a = (a_1, \ldots, a_n) \in D(s)$, $s_j, s'_j \in S_j$, $j = 1, \ldots, n$.

The following theorem states that the compound CMDP indeed models n independent projects. Furthermore, optimizing each project solely is equivalent to optimizing the compound CMDP as a whole.

Theorem 4.1.11. *Let* (CMDP_i), $i = 1, \ldots, n$, *be independent CMDPs such that* $\mathscr{K}_i \equiv 0$ *and* $\alpha_i = \alpha$. *For* $i = 1, \ldots, n$, *let* v_i^* *be the optimal value function of* (CMDP_i). *Then the compound CMDP* $(\mathrm{CMDP}) = (\mathrm{CMDP}_i)_{i=1,\ldots,n}$ *has the optimal value function* $v^*(s) = \sum_{i=1}^{n} v_i^*(s_i)$, *and a decision rule* $\mu^*(s) = (\mu_1^*(s_1), \ldots, \mu_n^*(s_n))$ *is optimal if and only if* μ_i^* *is optimal for* (CMDP_i) *for all* $i = 1, \ldots, n$.

Proof. For $i \in \{1, \ldots, n\}$, the optimal value function v_i^* of (CMDP_i) satisfies the Bellman equation

$$v_i^*(s_i) = \min_{a_i \in D_i(s_i)} \left\{ \frac{C_i(s_i, a_i)}{\lambda_i(s_i, a_i) + \alpha} + \frac{\lambda_i(s_i, a_i)}{\lambda_i(s_i, a_i) + \alpha} \sum_{s'_i \in S_i} p^{a_i}_{i,s_i,s'_i} v_i^*(s'_i) \right\}, \quad s_i \in S_i,$$

which is equivalent to

$$(\lambda_i(s_i, a_i) + \alpha) v_i^*(s_i) \leq C_i(s_i, a_i) + \lambda_i(s_i, a_i) \sum_{s'_i \in S_i} p^{a_i}_{i,s_i s'_i} v_i^*(s'_i) \quad ((s_i, a_i) \in D_i) \tag{4.8}$$

with equality for at least one $a_i \in D_i(s_i)$ for every $s_i \in S_i$. The value function v^* of (CMDP) satisfies the Bellman equation

$$v^*(s) = \min_{a \in D(s)} \left\{ \frac{C(s,a)}{\lambda(s,a) + \alpha} + \frac{\lambda(s,a)}{\lambda(s,a) + \alpha} \sum_{s' \in S} p^a_{ss'} v^*(s') \right\}, \quad s \in S,$$

or equivalently

$$(\lambda(s,a) + \alpha) v^*(s) \leq C(s,a) + \lambda(s,a) \sum_{s' \in S} p^a_{ss'} v^*(s') \quad ((s,a) \in D) \tag{4.9}$$

with equality for at least one $a \in D(s)$ for every $s \in S$. We show that $v^*(s) = \sum_{i=1} v_i^*(s_i)$, $s \in S$, satisfies condition (4.9). Therefore, let $(s,a) \in D$. Then

$$\begin{aligned}
&(\lambda(s,a) + \alpha) \sum_{i=1}^{n} v_i^*(s_i) \leq C(s,a) + \lambda(s,a) \sum_{s' \in S} p^a_{ss'} \sum_{i=1}^{n} v_i^*(s'_i) \\
\Leftrightarrow \quad &\left(\sum_{i=1}^{n} \lambda_i(s_i, a_i) + \alpha \right) \sum_{i=1}^{n} v_i^*(s_i) \leq \sum_{i=1}^{n} C_i(s_i, a_i) + \lambda(s,a) \sum_{i=1}^{n} \frac{\lambda_i(s_i, a_i)}{\lambda(s,a)} \sum_{s'_i \in S_i} p^{a_i}_{i,s_i s'_i} \left[v_i^*(s'_i) + \sum_{j \neq i} v_j^*(s_j) \right] \\
\Leftrightarrow \quad &\sum_{i=1}^{n} [(\lambda_i(s_i, a_i) + \alpha) v_i^*(s_i)] + \sum_{i=1}^{n} \sum_{j \neq i} \lambda_i(s_i, a_i) v_j^*(s_j) \\
&\qquad \leq \sum_{i=1}^{n} \left[C_i(s_i, a_i) + \lambda_i(s_i, a_i) \sum_{s'_i \in S_i} p^{a_i}_{i,s_i s'_i} v_i^*(s'_i) \right] + \sum_{i=1}^{n} \lambda_i(s_i, a_i) \sum_{j \neq i} v_j^*(s_j)
\end{aligned}$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} \left[(\lambda_i(s_i,a_i) + \alpha)\, v_i^*(s_i) \right] \leq \sum_{i=1}^{n} \left[C_i(s_i,a_i) + \lambda_i(s_i,a_i) \sum_{s_i' \in S_i} p_{i,s_i s_i'}^{a_i} v_i^*(s_i') \right]. \tag{4.10}$$

Inequality (4.10) is satisfied for all $(s,a) \in D$ in view of (4.8). Moreover, equality in (4.10) holds if and only if $a_i \in D(s_i)$ is optimal for (CMDP_i) for all $i \in \{1,\ldots,n\}$. Therefore, $\sum_{i=1}^n v_i^*$ is the optimal value function and $\mu^*(s) = (\mu_1^*(s_1),\ldots,\mu_n^*(s_n))$, $s \in S$, is an optimal decision rule if and only if μ_i^* is an optimal decision rule for (CMDP_i) for every $i = 1,\ldots,n$. □

So, it is sufficient to compute optimal policies for each sector separately. Since one has to consider only n CMDPs with respective restriction sets D_i, instead of solving one large CMDP with restriction set $\bigtimes_{i=1}^n D_i$, which reduces the dimensionality, the numerical complexity is reduced. By this theorem, the result from the preceding sections can be carried over to the case of an infrastructure which has no dependencies between its sectors and the decision maker is able to perform actions from the set A in all states.

Corollary 4.1.12. *Let the infrastructure consist of n sectors. Let $N \equiv 0$ and $r = n$. If for each sector Assumption 4.1.2 holds, then there is a decision rule $\mu^* = (\mu_1^*,\ldots,\mu_n^*)$ such that μ_i^* is a CTP for every $i \in \{1,\ldots,n\}$. Furthermore, action $(0,\ldots,0) \in A$ is optimal in state $(0,\ldots,0) \in S$.*

Proof. The result follows immediately from Theorems 4.1.3 and 4.1.11. □

But if dependencies come into play or if the parameters are more complicated than in Assumption 4.1.2, optimal decision rules are much more unstructured as could already be seen from Figure 3.4. Infrastructures with dependencies but again without resource restrictions are considered in the subsequent section. But strong results as the optimality of a CTP under certain assumptions cannot be derived in this case.

4.2. Coupled and Decoupled CMDPs

In this section, we consider a generalization of the threat model of chapter 3. As we have seen in the preceding section, an infrastructure with $N \equiv 0$ without resource restrictions can be seen as a compound CMDP. If the infrastructure has dependencies but if there are again no resource restrictions, then it can be seen as a coupled CMDP in the following sense.

Definition 4.2.1. Let $(\mathrm{CMDP}_i) = (S_i, A_i, D_i, \Lambda_i, P_i, \mathscr{K}_i, \mathscr{C}_i, \alpha_i)$, $i = 1, \ldots, n$, be CMDPs as defined in Definition 2.1.1 such that $\mathscr{K}_i \equiv 0$, $i = 1, \ldots, n$, and such that they have the same discount rate $\alpha_i = \alpha > 0$. We define the *coupled CMDP* $((\mathrm{CMDP}_i)_{i=1,\ldots,n}, P^{\mathrm{c}}, \kappa) := (S, A, D, \Lambda, P, \mathscr{C}, \mathscr{K}, \alpha)$ by

$$S := \bigtimes_{i=1}^{n} S_i, \quad A := \bigtimes_{i=1}^{n} A_i, \quad D(s_1,\ldots,s_n) := \bigtimes_{i=1}^{n} D(s_i), \quad \lambda((s_1,\ldots,s_n),(a_1,\ldots,a_n)) := \sum_{i=1}^{n} \lambda_i(s_i,a_i),$$

$$C((s_1,\ldots,s_n),(a_1,\ldots,a_n)) := \sum_{i=1}^{n} C_i(s_i,a_i), \quad \mathscr{K} \equiv 0, \quad (s_1,\ldots,s_n) \in S, \quad (a_1,\ldots,a_n) \in D(s_1,\ldots,s_n),$$

and P is given by the following mechanism: let $N \in \{0,1\}^{n \times n}$ be a matrix with $N(i,i) = 0$, $i = 1, \ldots, n$. Let $N(i) := \{j : N(i,j) = 1\}$ and

$$S_{N(i)} := \bigtimes_{j \in N(i)} S_j, \quad i = 1, \ldots, n.$$

Moreover, let $\kappa \in [0,1]$ be the *coupling mechanism probability* and let $p_{i,s_i s_i', s_{N(i)} s_{N(i)}'}^{\mathrm{c},a_i}$ be the *coupling transition probabilities*, $(s_i,a_i) \in D_i$, $s_i' \in S_i$, $s_{N(i)}$, $s_{N(i)}' \in S_{N(i)}$, $i = 1, \ldots, n$. After an $\mathrm{Exp}(\lambda(s,a))$-distributed time for $(s,a) \in D$, there is a transition in (CMDP_i) from s_i to s_i' with probability $\lambda_i(s_i,a_i)/\lambda((s_1,\ldots,s_n),(a_1,\ldots,a_n))\, p_{i,s_i s_i'}^{a_i}$. With probability κ, this transition causes the states of those (CMDP_j) with $N(i,j) = 1$ to transition from $s_{N(i)}$ to $s_{N(i)}'$ with probability $p_{i,s_i s_i', s_{N(i)} s_{N(i)}'}^{\mathrm{c},a_i}$ at the same time when action a_i was chosen in (CMDP_i). With probability $1-\kappa$, there are no state transitions in these (CMDP_j). If $N(i,j) = 0$, $i \neq j$, then there is no state transition in (CMDP_j).

Given a coupled CMDP $((\mathrm{CMDP}_i)_{i=1,\ldots,n}, P^{\mathrm{c}}, \kappa)$, we define the *associated decoupled CMDP* by the coupled CMDP $((\mathrm{CMDP}_i)_{i=1,\ldots,n}, P^{\mathrm{c}}, 0)$ (or equivalently by $((\mathrm{CMDP}_i)_{i=1,\ldots,n}, 0, \kappa)$ or by the compound CMDP $(\mathrm{CMDP}_i)_{i=1,\ldots,n}$) in which no coupling is present.

In this model, coupling arises in two ways: when there is a state transition in one of the (CMDP_i), then, in a first experiment, it is determined whether there is a coupling with probability κ. If the answer of the first experiment is that there is a coupling, then the transitions of those (CMDP_j) with $N(i,j)=1$ are determined by a second experiment with transition probabilities according to P^{c}.

Note that the threat model of chapter 3 is a coupled CMDP with $\kappa = 1$ if $r = |\Sigma|$. Each sector of the infrastructure is one of the (CMDP_i) forming the coupled CMDP. If a threat event occurs or an elementary action is completed in some sector $\sigma \in \Sigma$, then its threat level changes. Furthermore, dependent sectors change their respective threat levels with given transition functions depending on the current and the subsequent threat level of σ.

Our aim is to find bounds on the difference between the optimal value functions of the coupled model of Definition 4.2.1, denoted by $v^*_{\mathrm{c},\kappa}$, and the optimal value function of the associated decoupled model, denoted by v^*_{dc}. At first, we define the fixed point operators of the coupled CMDP and of the associated decoupled CMDP by

$$T_{\mathrm{c},\kappa}v(s) \coloneqq \min_{a\in D(s)}\left\{\frac{1}{\lambda(s,a)+\alpha}\left[C(s,a)+\sum_{i=1}^{n}\lambda_i(s_i,a_i)\sum_{s'_i\in S_i}p^{a_i}_{i,s_is'_i}\left[\kappa\sum_{s'_{N(i)}\in S_{N(i)}}p^{\mathrm{c},a_i}_{i,s_is'_i,s_{N(i)}s'_{N(i)}}\,v\left(s\leftarrow(i,s'_i,s'_{N(i)})\right)\right.\right.\right.$$
$$\left.\left.\left.+(1-\kappa)\,v(s_1,\ldots,s_{i-1},s'_i,s_{i+1},\ldots,s_n)\right]\right]\right\},$$

$$T_{\mathrm{dc}}v(s) \coloneqq \min_{a\in D(s)}\left\{\frac{1}{\lambda(s,a)+\alpha}\left[C(s,a)+\sum_{i=1}^{n}\lambda_i(s_i,a_i)\sum_{s'_i\in S_i}p^{a_i}_{i,s_is'_i}\,v(s_1,\ldots,s_{i-1},s'_i,s_{i+1},\ldots,s_n)\right]\right\},$$

$s=(s_1,\ldots,s_n)\in S$, $a=(a_1,\ldots,a_n)\in D(s)$, where $v:S\to\mathbb{R}$. For $s\in S$ and $i\in\{1,\ldots,n\}$, we define $\left(s\leftarrow(i,s'_i,s'_{N(i)})\right)\in S$ as the state in which the ith component of s is replaced with s'_i and the components of $N(i)$ are replaced with $s'_{N(i)}$. Of course, we have $v^*_{\mathrm{c},\kappa}(s)=T_{\mathrm{c},\kappa}v^*_{\mathrm{c},\kappa}(s)$ and $v^*_{\mathrm{dc}}(s)=T_{\mathrm{dc}}v^*_{\mathrm{dc}}(s)$ for all $s\in S$. We make use of the following definitions:

$$\gamma^* \coloneqq \max_{(s,a)\in D}\left\{\frac{\lambda(s,a)}{\lambda(s,a)+\alpha}\right\}$$

is the Lipschitz constant of both contraction mappings $T_{\mathrm{c},\kappa}$ and T_{dc}. Note that $0\le\gamma^*<1$ since D is finite. The optimal value functions of the underlying (CMDP_i), $i=1,\ldots,n$, are denoted by v^*_i. For a finite set $M\neq\emptyset$, we define the *span* of the function $v:M\to\mathbb{R}$ by

$$\mathrm{span}(v) \coloneqq \max_{x\in M}v(x)-\min_{x\in M}v(x).$$

At first, we have the following lemmas.

Lemma 4.2.2. *Let $M\neq\emptyset$ be a finite set, $\sum_{s\in M}p_s=1$, $p_s\ge 0$ $(s\in M)$ and $v:M\to\mathbb{R}$. For all $s\in M$, we have*

$$\left|v(s)-\sum_{s'\in M}p_{s'}\,v(s')\right|\le\mathrm{span}(v).$$

Proof. For $s\in M$, we compute

$$\left|v(s)-\sum_{s'\in M}p_{s'}\,v(s')\right|=\left|\sum_{s'\in M}p_{s'}\,\left[v(s)-v(s')\right]\right|\le\sum_{s'\in M}p_{s'}\,\left|v(s)-v(s')\right|\le\sum_{s'\in M}p_{s'}\,\mathrm{span}(v)=\mathrm{span}(v).$$ □

Lemma 4.2.3. *Let $((\mathrm{CMDP}_i)_{i=1,\ldots,n},P^{\mathrm{c}},\kappa)$ be a coupled CMDP and let $v:S\to\mathbb{R}$. Then*

$$\|T_{\mathrm{c},\kappa}v-T_{\mathrm{dc}}v\|_\infty\le\kappa\,\gamma^*\,\mathrm{span}(v).$$

Proof. We compute

$$\|T_{\mathrm{c},\kappa}v-T_{\mathrm{dc}}v\|_\infty$$
$$=\max_{s\in S}\left|\min_{a\in D(s)}\left\{\frac{1}{\lambda(s,a)+\alpha}\left[C(s,a)+\sum_{i=1}^{n}\lambda_i(s_i,a_i)\sum_{s'_i\in S_i}p^{a_i}_{i,s_is'_i}\left[\kappa\sum_{s'_{N(i)}\in S_{N(i)}}p^{\mathrm{c},a_i}_{i,s_is'_i,s_{N(i)}s'_{N(i)}}\,v\left(s\leftarrow(i,s'_i,s'_{N(i)})\right)\right.\right.\right.\right.$$

$$
\begin{aligned}
&\qquad\qquad + (1-\kappa)\, v(s_1,\dots,s_{i-1},s_i',s_{i+1},\dots,s_n)\Bigg]\Bigg]\Bigg\} \\
&\qquad - \min_{a\in D(s)}\left\{\frac{1}{\lambda(s,a)+\alpha}\left[C(s,a)+\sum_{i=1}^{n}\lambda_i(s_i,a_i)\sum_{s_i'\in S_i}p_{i,s_is_i'}^{a_i}\, v(s_1,\dots,s_{i-1},s_i',s_{i+1},\dots,s_n)\right]\right\}\Bigg| \\
&\overset{\text{Lemma A.5}}{\leq} \max_{(s,a)\in D}\left\{\frac{1}{\lambda(s,a)+\alpha}\left[\sum_{i=1}^{n}\lambda_i(s_i,a_i)\sum_{s_i'\in S_i}p_{i,s_is_i'}^{a_i}\left|\kappa\sum_{s'_{N(i)}\in S_{N(i)}}p_{i,s_is_i',s_{N(i)}s'_{N(i)}}^{\mathrm{c},a_i}\, v\left(s\leftarrow(i,s_i',s'_{N(i)})\right)\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. - \kappa\, v(s_1,\dots,s_{i-1},s_i',s_{i+1},\dots,s_n)\right|\right]\right\} \qquad (4.11) \\
&\overset{\text{Lemma 4.2.2}}{\leq} \kappa\gamma^*\operatorname{span}(v). \qquad \square
\end{aligned}
$$

Thus, the functions $T_{\mathrm{c},\kappa}v$ converge uniformly to $T_{\mathrm{dc}}v$ as $\kappa\to 0$, i. e., as the coupling mechanism vanishes. Now, we are ready to compute the following bound.

Proposition 4.2.4. *Let* $((\mathrm{CMDP}_i)_{i=1,\dots,n}, P^{\mathrm{c}}, \kappa)$ *be a coupled CMDP. Then it holds*

$$
\left\|v_{\mathrm{c},\kappa}^* - v_{\mathrm{dc}}^*\right\|_\infty \leq \frac{\kappa\gamma^*}{1-\gamma^*}\max_{i=1,\dots,n}\left\{\sum_{j\in N(i)}\operatorname{span}(v_j^*)\right\}.
$$

Proof. By the triangle inequality, we have

$$
\begin{aligned}
\left\|v_{\mathrm{c},\kappa}^* - v_{\mathrm{dc}}^*\right\|_\infty &= \left\|T_{\mathrm{c},\kappa}v_{\mathrm{c},\kappa}^* - T_{\mathrm{dc}}v_{\mathrm{dc}}^*\right\|_\infty \leq \left\|T_{\mathrm{c},\kappa}v_{\mathrm{c},\kappa}^* - T_{\mathrm{c},\kappa}v_{\mathrm{dc}}^*\right\|_\infty + \left\|T_{\mathrm{c},\kappa}v_{\mathrm{dc}}^* - T_{\mathrm{dc}}v_{\mathrm{dc}}^*\right\|_\infty \\
&\overset{\text{Lemma 2.3.3}}{\leq} \gamma^*\left\|v_{\mathrm{c},\kappa}^* - v_{\mathrm{dc}}^*\right\|_\infty + \left\|T_{\mathrm{c},\kappa}v_{\mathrm{dc}}^* - T_{\mathrm{dc}}v_{\mathrm{dc}}^*\right\|_\infty,
\end{aligned}
$$

from which we obtain $\left\|v_{\mathrm{c},\kappa}^* - v_{\mathrm{dc}}^*\right\|_\infty \leq \left\|T_{\mathrm{c},\kappa}v_{\mathrm{dc}}^* - T_{\mathrm{dc}}v_{\mathrm{dc}}^*\right\|_\infty/(1-\gamma^*)$. Since the associated decoupled CMDP is a compound CMDP, we have $v_{\mathrm{dc}}^*(s_1,\dots,s_n) = \sum_{i=1}^n v_i^*(s_i)$, $(s_1,\dots,s_n)\in S$, by Theorem 4.1.11. For $(s,a)\in D$, $s_i'\in S_i$, $i\in\{1,\dots,n\}$, we have

$$
\begin{aligned}
&\left|\sum_{s'_{N(i)}\in S_{N(i)}}p_{i,s_is_i',s_{N(i)}s'_{N(i)}}^{\mathrm{c},a_i}\, v_{\mathrm{dc}}^*\left(s\leftarrow(i,s_i',s'_{N(i)})\right) - v_{\mathrm{dc}}^*(s_1,\dots,s_{i-1},s_i',s_{i+1},\dots,s_n)\right| \\
&\quad\leq \sum_{j\in N(i)}\left|\sum_{s'_{N(i)}\in S_{N(i)}}p_{i,s_is_i',s_{N(i)}s'_{N(i)}}^{\mathrm{c},a_i}\, v_j^*\left(\left(s\leftarrow(i,s_i',s'_{N(i)})\right)_j\right) - v_j^*(s_j)\right| \overset{\text{Lemma 4.2.2}}{\leq} \sum_{j\in N(i)}\operatorname{span}(v_j^*) \qquad (4.12)
\end{aligned}
$$

since $v_j^*\left(\left(s\leftarrow(i,s_i',s'_{N(i)})\right)_j\right) = v_j^*(s_j)$ if $j\notin N(i)$. From (4.11) and (4.12), we obtain

$$
\left\|T_{\mathrm{c},\kappa}v_{\mathrm{dc}}^* - T_{\mathrm{dc}}v_{\mathrm{dc}}^*\right\|_\infty \leq \kappa\gamma^*\max_{i=1,\dots,n}\left\{\sum_{j\in N(i)}\operatorname{span}(v_j^*)\right\},
$$

from which the assertion follows. $\square$

The bound from Proposition 4.2.4 is easy to obtain because only γ^* and the solutions v_i^* of the underlying (CMDP_i), $i=1,\dots,n$, have to be known. Note again that it is much easier to derive the optimal value functions of the (CMDP_i) than the value function of the coupled CMDP $((\mathrm{CMDP}_i)_{i=1,\dots,n}, P^{\mathrm{c}}, \kappa)$ since the dimensions of the restriction sets are reduced. The bound is sharp in the sense that if the coupled model is indeed decoupled, i. e., $\kappa = 0$ or $N\equiv 0$, then it follows from Proposition 4.2.4 that $v_{\mathrm{c},\kappa}^* = v_{\mathrm{dc}}^*$. The bound might not be very good since γ^* is almost one if α is sufficiently small. But if the bound is not too large, one is able to obtain the approximate magnitude of the optimal value function of the coupled CMDP.

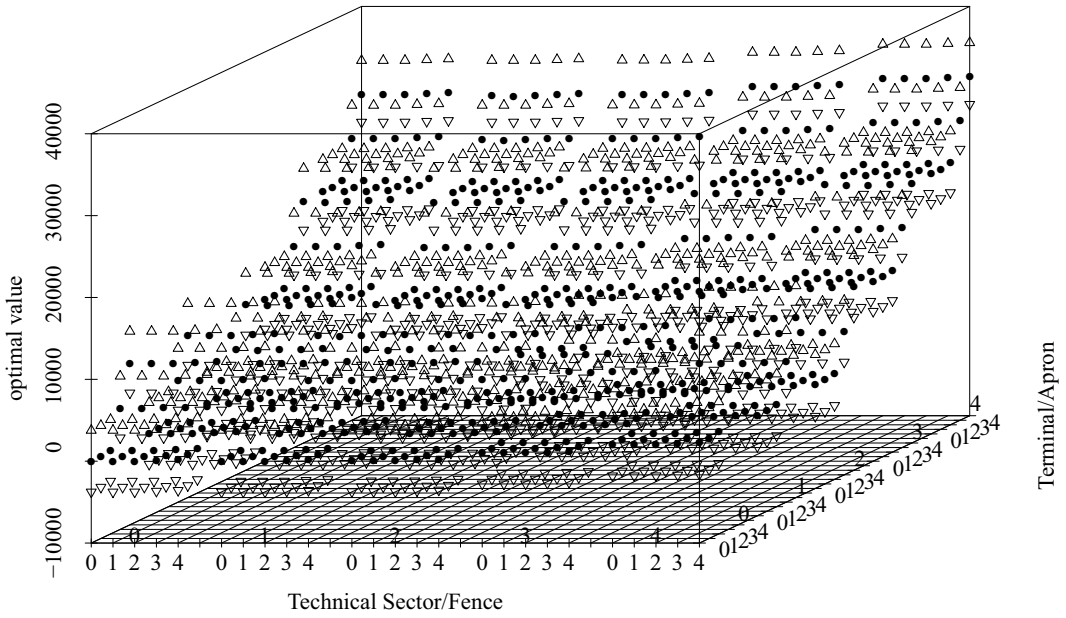


Figure 4.2.: Bounds on the optimal value function according to Proposition 4.2.4.

Example 4.2.5. Again, we consider the accompanying example of chapter 3. If $\alpha = \log(2)/24$, then $\gamma^*/(1-\gamma^*) \cdot \sum_{j\in N(j)} \operatorname{span}(v_j^*) \approx 10^{10}$ so that the bound of Proposition 4.2.4 is not good at all since the optimal value functions of the four sectors all satisfy $0 \leq v_\sigma^* \leq 2 \cdot 10^6$ which is four orders of magnitude lower. Therefore, we consider the same infrastructure with $\alpha = 1000$. Then we have $\gamma^*/(1-\gamma^*) \sum_{j\in N(j)} \operatorname{span}(v_j^*) \approx 3763$ which is quite a useful bound. In Figure 4.2, the optimal value function is illustrated together with the lower and upper bounds derived in Proposition 4.2.4. In the following Figures 4.3 and 4.4, cuts of Figure 4.2 through the states with fixed $s(\mathrm{T})$, $s(\mathrm{A})$ and with fixed $s(\mathrm{TS})$, $s(\mathrm{F})$ are illustrated. The lower bound is shown by a triangle with a downward pointing vertex and the upper bound is shown by a triangle with an upward pointing vertex. The bounds give a quite good approximation of the value function of the infrastructure.

Next, we investigate how the decision rule μ_{dc}^* which is optimal for the associated decoupled CMDP performs in the coupled model. This is of interest, since the derivation of μ_{dc}^* might be quite simple since it is given by optimal decision rules for the underlying (CMDP_i), $i = 1, \ldots, n$, by Theorem 4.1.11.

Proposition 4.2.6. *Let $((\mathrm{CMDP}_i)_{i=1,\ldots,n}, P^{\mathrm{c}}, \kappa)$ be a coupled CMDP. Let μ_{dc}^* be an optimal decision rule for the associated decoupled CMDP defined by $\mu_{\mathrm{dc}}^*(s_1,\ldots,s_n) := (\mu_1^*(s_1),\ldots,\mu_n^*(s_n))$, $(s_1,\ldots,s_n) \in S$, where μ_i^* is an optimal decision rule of (CMDP_i), $i = 1, \ldots, n$, and $v_{\mathrm{c},\kappa}^{\mu_{\mathrm{dc}}^*}$ be the value function of μ_{dc}^* for the coupled CMDP. Then*

$$\begin{aligned}&\left\| v_{\mathrm{c},\kappa}^{\mu_{\mathrm{dc}}^*} - v_{\mathrm{c},\kappa}^* \right\|_\infty \\ &\quad \leq \frac{\kappa}{1-\gamma^{\mu_{\mathrm{dc}}^*}} \left[\gamma^{\mu_{\mathrm{dc}}^*} \frac{\max_{(s,a)\in D} C(s,a) - \min_{(s,a)\in D} C(s,a)}{\alpha} + \frac{(1+\gamma^{\mu_{\mathrm{dc}}^*})\,\gamma^*}{1-\gamma^*} \max_{i=1,\ldots,n} \left\{ \sum_{j\in N(i)} \operatorname{span}(v_j^*) \right\} \right],\end{aligned}$$

where $\gamma^{\mu_{\mathrm{dc}}^*} := \max_{s\in S} \left\{ \lambda(s,\mu_{\mathrm{dc}}^*(s)) / (\lambda(s,\mu_{\mathrm{dc}}^*(s)) + \alpha) \right\}$.

Proof. For $v : S \to \mathbb{R}$, define the operators

$$T_{\mathrm{c},\kappa,\mu_{\mathrm{dc}}^*} v(s)$$

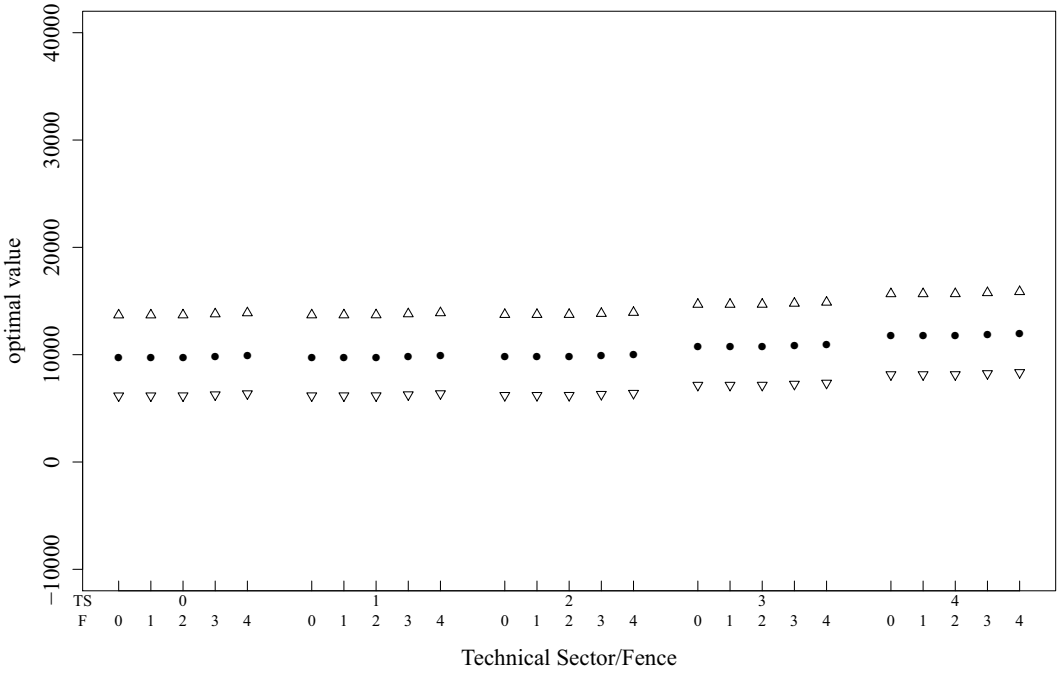


Figure 4.3.: Cut through the states with $s(\mathrm{T}) = 3$ and $s(\mathrm{A}) = 1$.

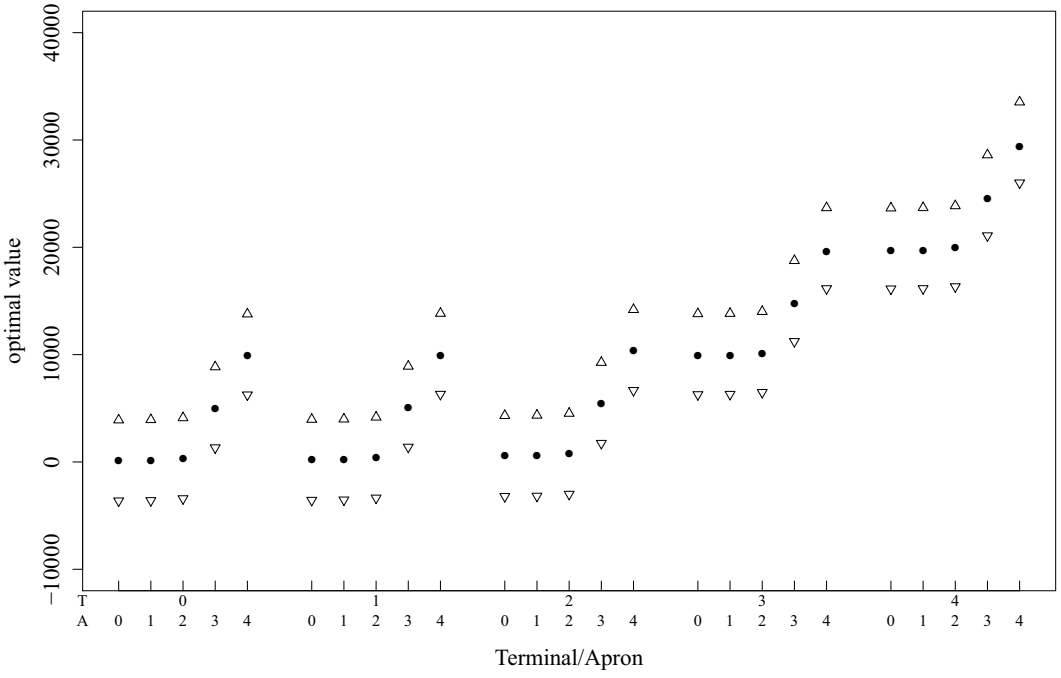


Figure 4.4.: Cut through the states with $s(\mathrm{TS}) = 2$ and $s(\mathrm{F}) = 3$.

$$\coloneqq \frac{1}{\lambda(s,\mu^*_{\mathrm{dc}}(s))+\alpha}\left[C(s,\mu^*_{\mathrm{dc}}(s))+\sum_{i=1}^{n}\lambda_i(s_i,\mu_i^*(s_i))\sum_{s_i'\in S_i}p_{i,s_is_i'}^{\mu_i^*(s_i)}\left[\kappa\sum_{s'_{N(i)}\in S_{N(i)}}p_{i,s_is_i',s_{N(i)}s'_{N(i)}}^{\mathrm{c},\mu^*(s_i)}\,v\left(s\leftarrow(i,s_i',s'_{N(i)})\right)\right.\right.$$
$$\left.\left.+(1-\kappa)\,v(s_1,\dots,s_{i-1},s_i',s_{i+1},\dots,s_n)\right]\right],$$

$$T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v(s)\coloneqq\frac{1}{\lambda(s,\mu^*_{\mathrm{dc}}(s))+\alpha}\left[C(s,\mu^*_{\mathrm{dc}}(s))+\sum_{i=1}^{n}\lambda_i(s_i,\mu_i^*(s_i))\sum_{s_i'\in S_i}p_{i,s_is_i'}^{\mu_i^*(s_i)}\,v(s_1,\dots,s_{i-1},s_i',s_{i+1},\dots,s_n)\right],\quad s\in S.$$

Then we have $T_{\mathrm{c},\kappa,\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}=v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}$ and $T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v^*_{\mathrm{dc}}=v^*_{\mathrm{dc}}$. By the triangle inequality, we have

$$\left\|v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-v^*_{\mathrm{c},\kappa}\right\|_\infty\le\left\|v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}\right\|_\infty+\left\|T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-v^*_{\mathrm{dc}}\right\|_\infty+\left\|v^*_{\mathrm{dc}}-v^*_{\mathrm{c},\kappa}\right\|_\infty.\tag{4.13}$$

For the first summand of (4.13), we obtain

$$\begin{aligned}\left\|v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}\right\|_\infty&=\left\|T_{\mathrm{c},\kappa,\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}\right\|_\infty\overset{\text{Lemma 4.2.3}}{\le}\kappa\,\gamma^{\mu^*_{\mathrm{dc}}}\operatorname{span}(v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}})\\&\le\kappa\,\gamma^{\mu^*_{\mathrm{dc}}}\frac{\max_{(s,a)\in D}C(s,a)-\min_{(s,a)\in D}C(s,a)}{\alpha},\end{aligned}\tag{4.14}$$

where we set $D(s)$ to $\{\mu^*_{\mathrm{dc}}(s)\}$ for all $s\in S$ when applying Lemma 4.2.3. We also made use of the trivial bounds $\min_{(s',a')\in D}C(s',a')/\alpha\le v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}(s)\le\max_{(s',a')\in D}C(s',a')/\alpha$ for all $s\in S$. The second summand of (4.13) is bounded by

$$\begin{aligned}\left\|T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-v^*_{\mathrm{dc}}\right\|_\infty&=\left\|T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-T_{\mathrm{dc},\mu^*_{\mathrm{dc}}}v^*_{\mathrm{dc}}\right\|_\infty\le\gamma^{\mu^*_{\mathrm{dc}}}\left\|v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-v^*_{\mathrm{dc}}\right\|_\infty\\&\le\gamma^{\mu^*_{\mathrm{dc}}}\left(\left\|v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-v^*_{\mathrm{c},\kappa}\right\|_\infty+\left\|v^*_{\mathrm{c},\kappa}-v^*_{\mathrm{dc}}\right\|_\infty\right).\end{aligned}\tag{4.15}$$

The third summand of (4.13) is considered in Proposition 4.2.4. So, collecting the results of (4.14), (4.15) and Proposition 4.2.4, we finally obtain

$$\begin{aligned}&\left\|v_{\mathrm{c},\kappa}^{\mu^*_{\mathrm{dc}}}-v^*_{\mathrm{c},\kappa}\right\|_\infty\\&\le\frac{1}{1-\gamma^{\mu^*_{\mathrm{dc}}}}\left[\kappa\,\gamma^{\mu^*_{\mathrm{dc}}}\frac{\max_{(s,a)\in D}C(s,a)-\min_{(s,a)\in D}C(s,a)}{\alpha}+\gamma^{\mu^*_{\mathrm{dc}}}\frac{\kappa\,\gamma^*}{1-\gamma^*}\max_{i=1,\dots,n}\left\{\sum_{j\in N(i)}\operatorname{span}(v_j^*)\right\}\right.\\&\qquad\left.+\frac{\kappa\,\gamma^*}{1-\gamma^*}\max_{i=1,\dots,n}\left\{\sum_{j\in N(i)}\operatorname{span}(v_j^*)\right\}\right]\\&=\frac{\kappa}{1-\gamma^{\mu^*_{\mathrm{dc}}}}\left[\gamma^{\mu^*_{\mathrm{dc}}}\frac{\max_{(s,a)\in D}C(s,a)-\min_{(s,a)\in D}C(s,a)}{\alpha}+\frac{(1+\gamma^{\mu^*_{\mathrm{dc}}})\,\gamma^*}{1-\gamma^*}\max_{i=1,\dots,n}\left\{\sum_{j\in N(i)}\operatorname{span}(v_j^*)\right\}\right].\end{aligned}$$

□

Remark 4.2.7. A necessary condition so that Proposition 4.2.6 gives a nontrivial bound is that $\kappa\,\gamma^{\mu^*_{\mathrm{dc}}}/(1-\gamma^{\mu^*_{\mathrm{dc}}})<1\Leftrightarrow\gamma^{\mu^*_{\mathrm{dc}}}<1/(1+\kappa)$ since the value function of any policy lies between $\min_{(s,a)\in D}C(s,a)/\alpha$ and $\max_{(s,a)\in D}C(s,a)/\alpha$.

If the coupling mechanism is sufficiently small, then μ^*_{dc} is an ε-optimal decision rule for the coupled CMDP. Unfortunately, the bound given in Proposition 4.2.6 might be very large since $\gamma^{\mu^*_{\mathrm{dc}}}\approx 1$ and $\gamma^*\approx 1$ for some small $\alpha>0$. This might be the case, e. g., when considering the threat model in which the coupling mechanism applies with probability $\kappa=1$.

Example 4.2.8. Let us again consider the accompanying example of chapter 3 with discount rate $\alpha=1000$. Then $\gamma^{\mu^*_{\mathrm{dc}}}\approx 0.045<1/2$. Then the bound of Proposition 4.2.6 is about 5150. An optimal policy for the associated decoupled CMDP is to perform elementary action 2 in every sector in every state. The maximal difference between the value function of this policy and the optimal value function of the surveillance task is about 248 which lies well below the calculated bound. Unfortunately, the policy which chooses elementary action 0 in every state also satisfies this bound in this example. This indicates that the bound might not be very good indeed at least for the surveillance task. Better results might be obtained for other models of coupled CMDPs.

5. Approximate Solution Methods

In this chapter, we consider the surveillance task with security staff size r of general parametrizations of infrastructures according to chapter 3. Numerical experiments have shown that it is impossible to compute an optimal decision rule with respect to the expected total discounted cost criterion for even quite small infrastructures. An infrastructure consisting of five sectors with three threat events for each sector, $g_{\max} = 4$ and $|A_0| = 3$ with staff size $r = 2$ is tractable through the linear programming method of Theorem 2.3.8. But the derivation of a solution of a similar infrastructure with six sectors is impossible through linear programming. In this case, the associated linear program could not be solved by the commercial solver CPLEX within 48 hours on a computer with an AMD Athlon 64 X2 Dual Core Processor 4600+ 2.41 GHz with 2 GB RAM with a Windows XP system running.

For a general infrastructure, the associated linear program for the surveillance task is not sparse. This is due to the transition functions of the threat events and of the elementary actions. The more threat events or elementary actions are modelled, the more states have non-zero probability to be reached at the next step so that the matrix of the linear program is becoming more dense. So, no efficient sparse matrix algorithms can be used for solving the surveillance task.

The fundamental cause why solutions are hard to obtain for large infrastructures is the so-called *curse of dimensionality*. As the size of the infrastructure increases, the state space increases exponentially and so does the restriction set if $r = |\Sigma|$. For fixed $r < |\Sigma|$, the growth of the restriction set is only polynomial. In the first case, there are $5^5 = 3{,}125$ states, whereas in the second case, there are $5^6 = 15{,}635$ states. From Proposition 3.3.9, we have $1 + 10 + 24 = 35$ actions in the first case and $1 + 12 + 40 = 53$ actions in the second case if $r = 2$. So, the linear program for the first infrastructure has 3,125 variables and 109,375 constraints, and the linear program for the second infrastructure has 15,635 variables and 828,655 constraints.

For Howard's policy improvement algorithm, one has to compute $T_a v(s)$ for all $(s,a) \in D$ for appropriate $v : S \to \mathbb{R}$. For large state spaces, this is impossible. Furthermore, one has to obtain the value function of a given decision rule, which consists of solving a system of $|S|$ linear equations by Remark 2.3.5. Again, this is impossible if the state space is large.

So, exact determination of an optimal policy is not possible. But we try to find a decision rule which is almost as good as the optimal decision rule with the help of approximation techniques.

There are several approximation methods for MDPs considered in the literature. A very good overview on general approximate dynamic programming techniques is given in (Powell, 2007). Numerous simulation-based solution methods are considered in (Chang et al., 2007).

5.1. Requirements on Approximate Solution Methods

The aim is to find an approximate solution method for the surveillance task such that a resulting policy can be used as the basis of a decision support system. Since in the surveillance task, decisions have to be made under time pressure in critical situations, an important requirement is that an appropriate action should be accessible within a short time. Therefore, computationally intensive parts of a solution method should be computable in advance.

Another important point is that the essential parts of the method, from which the policy is obtained, should be able to be stored on a hard disk drive. As an example, consider an infrastructure such that $|\Sigma| = 12$ and $g_{\max} = 4$. Assume that a decision rule is stored in a naive way as a lookup table as shown below where each row stands for one state, and each symbol for an elementary action for the respective sector:

```
      ⋮
000000000000
100010000000
002000001000
      ⋮
```

If each symbol requires 8 bits, then the whole policy requires $5^{12} \cdot 12 \cdot 8$ bits ≈ 2.9 GB of memory. The memory requirements grow exponentially with the number of sectors which makes it impossible to store a policy for large infrastructures in this manner. An infrastructure of size 15 requires about 450 GB of memory, and an infrastructure of size 20 requires about 1.9 PB.

5.2. Value Iteration

In section 2.3.1, we briefly introduced the value iteration method for approximately solving the expected total discounted cost criterion. Essential for this algorithm is the execution of a one-step improvement by updating $v_{n+1} \coloneqq Tv_n$, $n \in \mathbb{N}_0$. As for Howard's policy improvement algorithm, this includes computing $T_a v_n(s)$ for all $(s,a) \in D$, in which all necessary parameters have to be determined from the parametrization. If S is sufficiently large, even one minimization step is impossible to accomplish. For example, if $|\Sigma| = 12$, $g_{\max} = 4$, $|A_0| = 3$ and $r = 2$, one has to consider $5^{12} = 244{,}140{,}625$ states and $1 + 24 + 264 = 289$ different actions for each state. Furthermore, the convergence of the value iteration is linear at rate $\gamma^* = \max_{(s,a)\in D}\{\lambda(s,a)/(\lambda(s,a)+\alpha)\}$, which is very slow since γ^* might be almost one if $\alpha > 0$ is small.

5.3. Approximate Linear Programming for the Surveillance Task

In this section, we want to solve the problem by approximating the optimal value by a linear combination of basis functions. This is similar to (Guestrin et al., 2003) where factored MDPs in discrete time are considered. In a factored MDP, the state space S consists of several components such that $S = \bigtimes_{i=1}^{n} S_i$ for some $n \in \mathbb{N}$. The state transition of one component depends on the current states of only few other components and on the current action. At every time step, each component changes its state independently of the simultaneous transitions of the other components. Often, the cost function is also factored, meaning that it is a sum of cost functions each representing one component and the action only. Then there are efficient approximate algorithms for solving the factored MDP. (Kan and Shelton, 2008) carry over this approach to factored CMDPs. A state transition may occur in each component of the state space independently of the state transitions of the other components where the state transitions only depend on the current state of some components. Note that only one component changes its state at every transition time in a factored CMDP since the sojourn times of the states of the components are independent and exponentially distributed.

The threat model is almost a factored CMDP in the sense of (Kan and Shelton, 2008). Threat events occur and elementary actions are accomplished, which both have influence on the affected sector only. But due to the dependency structure, dependent sectors are influenced and may change their threat levels, too. This is the difference to factored CMDPs.

5.3.1. General Approximate Linear Programming

We begin with a general method for approximately solving CMDPs. In the discrete-time case, it is considered, e. g., in (de Farias and van Roy, 2003) and in (Powell, 2007), section 9.2.5. But in continuous time, the formulation is straightforward. In this approach, the optimal value function v^* is approximated by a linear combination $\hat{v}$ of given basis functions h_1, $\ldots$, $h_m : S \to \mathbb{R}$ such that $\hat{v}(s) = \sum_{i=1}^{m} w_i^* h_i(s)$, $s \in S$, for some $w_i^* \in \mathbb{R}$, $i = 1, \ldots, m$. Instead of using the exact linear programming formulation of Theorem 2.3.8, we use its approximate version by substituting $v(s)$ with its approximate counterpart $\sum_{i=1}^{m} w_i h_i(s)$, $s \in S$, where $\zeta : S \to \mathbb{R}_{>0}$ is given:

$$\begin{aligned} &\text{Maximize} \sum_{s\in S} \zeta(s) \sum_{i=1}^{m} w_i h_i(s) \\ &\text{under the constraint} \qquad\qquad (\mathrm{LP}_{\mathrm{appr}}) \\ &\sum_{i=1}^{m} w_i h_i(s) - \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s'\in S} p_{ss'}^{a} \sum_{i=1}^{m} w_i h_i(s') \le c(s,a) \quad ((s,a) \in D). \end{aligned}$$

In ($\mathrm{LP}_{\mathrm{appr}}$), the variables are the $w_1, \ldots, w_m$. In contrast to the exact linear programming formulation, the choice of the state-relevance weights ζ influences the solutions $w_1^*, \ldots, w_m^*$ as stated in (de Farias and van Roy, 2003). Whereas an optimal decision rule can be obtained directly from the solution of the respective exact linear program, there is no such way in approximate linear programming. In general, there may not be an action $a \in D(s)$ such that the constraint corresponding to (s,a) is active for every $s \in S$. Nevertheless, an approximate decision rule $\hat{\mu}$ may be derived from the *greedy algorithm,* meaning that $\hat{\mu}(s)$ is chosen such that

$$\hat{\mu}(s) \in \operatorname*{arg\,min}_{a\in D(s)} \left\{ c(s,a) + \frac{\lambda(s,a)}{\lambda(s,a)+\alpha} \sum_{s'\in S} p_{ss'}^{a} \hat{v}(s') \right\}, \quad s \in S. \tag{5.1}$$

The decision rule $\hat{\mu}$ leads to the *greedy policy with respect to* $\hat{v}$. In (Bertsekas, 2001), pp. 55–58, this method is called the *one-step look-ahead policy with respect to* $\hat{v}$. Note that the greedy algorithm is a one-step iteration of the value iteration, which we wanted to avoid due to its complexity. Furthermore, (de Farias and van Roy, 2003) give bounds for the

difference of the optimal value function and the approximate value function with respect to the supremum norm and the weighted supremum norm. Continuous-time versions of these bounds can be found in (Kan and Shelton, 2008). Moreover, (de Farias and van Roy, 2003) also make numerical studies considering several queueing networks. They show that with a good choice of the basis functions and of the state-relevance weights, the derived greedy policy is almost optimal.

5.3.2. Structured CMDPs

We make use of the approximate linear programming approach of section 5.3.1. The goal is to reduce the number of constraints of ($\mathrm{LP}_{\mathrm{appr}}$) by exploiting the special structure of the threat model. Consider the parametrization of an infrastructure according to chapter 3. We assume that the security staff size is r. For every $\sigma \in \Sigma$, let $\Sigma(\sigma) \coloneqq \{\sigma^* \in \Sigma : N(\sigma,\sigma^*) = 1\} \cup \{\sigma\}$ be the set of all sectors that are affected by a threat event occurring in σ or by an accomplished elementary action in σ. Following (Kan and Shelton, 2008), the constraint of ($\mathrm{LP}_{\mathrm{appr}}$) for the respective surveillance task holds if and only if

$$
\begin{aligned}
&\min_{\substack{s\in S:\\ a\in D(s)}} \left\{ C(s,a) + \sum_{i=1}^{m} w_i \left[-(\lambda(s,a)+\alpha)\, h_i(s) + \lambda(s,a) \sum_{s'\in S} p^a_{ss'} h_i(s') \right] \right\} \geq 0 \quad (a \in A) \\
\Leftrightarrow \quad &\min_{\substack{s\in S:\\ a\in D(s)}} \left\{ C(s,a) + \sum_{i=1}^{m} w_i \left[-\left(\alpha - \sum_{\sigma\in I(i)} \lambda_\sigma(s(\sigma),a(\sigma)) \right) h_i(s) - \sum_{\sigma\notin I(i)} \lambda_\sigma(s(\sigma),a(\sigma))\, h_i(s) \right.\right. \\
&\qquad + \sum_{\sigma\in I(i)} \left[\lambda_{a(\sigma)}(\sigma) \sum_{g'\in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}(g')\, h_i\left(\zeta^{g'}_{a(\sigma),\sigma}(s)\right) + \sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma))\, h_i(\zeta_e(s)) \right] \\
&\qquad \left.\left. + \sum_{\sigma\notin I(i)} \left[\lambda_{a(\sigma)}(\sigma) \sum_{g'\in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}(g')\, h_i\left(\zeta^{g'}_{a(\sigma),\sigma}(s)\right) + \sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma))\, h_i(\zeta_e(s)) \right] \right] \right\} \geq 0 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (a \in A),
\end{aligned}
\tag{5.2}
$$

where $\lambda_\sigma(s(\sigma),a(\sigma)) \coloneqq \sum_{e\in\mathscr{E}(\sigma)} \lambda_e(\sigma) + \lambda_{a(\sigma)}(\sigma)$, $\sigma \in \Sigma$, the $I(i)$, $i = 1,\dots,m$, are arbitrary subsets of Σ and ζ_e and $\zeta^{g'}_{a(\sigma),\sigma}$ are the functions which determine the subsequent states as introduced in sections 3.3.3 and 3.3.4. If for some $a \in A$, there is no $s \in S$ such that $a \in D(s)$, we define the corresponding constraint of (5.2) to be valid. Now, we seek for subsets $I(i)$ such that the computation of (5.2) is simplified. To this end, let $\tilde{\Sigma}(\sigma) \coloneqq \{\sigma^* \in \Sigma : N(\sigma^*,\sigma) = 1\} \cup \{\sigma\}$, $\sigma \in \Sigma$, be the set of all sectors that have influence on the threat level of σ. Consider the basis function $h_i : S \to \mathbb{R}$ for some $i \in \{1,\dots,m\}$. We assume that h_i only depends on the threat level of the sectors lying in the set $\Sigma_{h_i} \subset \Sigma$, i. e., $h_i(s') = h_i(s)$ for all s, $s' \in S$ such that $s(\sigma) = s'(\sigma)$ for all $\sigma \in \Sigma_{h_i}$. At last, we define $I(i)$ as the set of sectors that have influence on Σ_{h_i}, i. e., $I(i) \coloneqq \bigcup_{\sigma\in\Sigma_{h_i}} \tilde{\Sigma}(\sigma)$. The following lemma shows that this definition $I(i)$ is the right choice to simplify the constraint (5.2).

Lemma 5.3.1. *For every $(s,a) \in D$, we have*

$$
\begin{aligned}
&\sum_{\sigma\notin I(i)} \left[\lambda_{a(\sigma)}(\sigma) \sum_{g'\in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}(g')\, h_i\left(\zeta^{g'}_{a(\sigma),\sigma}(s)\right) + \sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma))\, h_i(\zeta_e(s)) \right] \\
&\quad = \sum_{\sigma\notin I(i)} \lambda(s(\sigma),a(\sigma))\, h_i(s), \quad i = 1,\dots,m.
\end{aligned}
\tag{5.3}
$$

Furthermore, constraint (5.2) *holds if and only if*

$$
\begin{aligned}
&\min_{\substack{s\in S:\\ a\in D(s)}} \left\{ C(s,a) + \sum_{i=1}^{m} w_i \left[-\left(\alpha - \sum_{\sigma\in I(i)} \lambda_\sigma(s(\sigma),a(\sigma)) \right) h_i(s) \right.\right. \\
&\qquad \left.\left. + \sum_{\sigma\in I(i)} \left[\lambda_{a(\sigma)}(\sigma) \sum_{g'\in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}(g')\, h_i\left(\zeta^{g'}_{a(\sigma),\sigma}(s)\right) + \sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma))\, h_i(\zeta_e(s)) \right] \right] \right\} \geq 0 \quad (a \in A).
\end{aligned}
\tag{5.4}
$$

Proof. We compute

$$
\sum_{\sigma\notin I(i)} \left[\lambda_{a(\sigma)}(\sigma) \sum_{g'\in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}(g')\, h_i\left(\zeta^{g'}_{a(\sigma),\sigma}(s)\right) + \sum_{e\in\mathscr{E}(\sigma)} \lambda_e(s(\sigma))\, h_i(\zeta_e(s)) \right]
$$

$$= \sum_{\sigma \notin I(i)} \left[\lambda_{a(\sigma)}(\sigma) \sum_{g' \in G} \Phi^{s(\sigma)}_{a(\sigma),\sigma}(g') h_i(s) + \sum_{e \in \mathscr{E}(\sigma)} \lambda_e(s(\sigma)) h_i(s) \right] = \sum_{\sigma \notin I(i)} \lambda(s(\sigma), a(\sigma)) h_i(s), \quad i = 1, \ldots, m,$$

for every $(s,a) \in D$. Substituting (5.3) into (5.2), yields the second assertion. □

Unfortunately, (5.4) is not a linear constraint. Nevertheless, it can be transformed into a linear constraint by a trick that also finds use in (Guestrin et al., 2003; Kan and Shelton, 2008). It can best be seen from an example since the formulation of a general proposition is rather tedious. Consider some function $f : M_1 \times M_2 \times M_3 \to \mathbb{R}$ such that $f(x_1,x_2,x_3) = f_1(x_1,x_2) + f_2(x_2,x_3) + f_3(x_3)$ where M_1, M_2, M_3 are non-empty finite sets. Then we can write

$$\min_{x_i \in M_i, i=1,2,3} \{f(x_1,x_2,x_3)\} = \min_{x_3 \in M_3} \left\{ f_3(x_3) + \min_{x_2 \in M_2} \left\{ f_2(x_2,x_3) + \min_{x_1 \in M_1} \{f_1(x_1,x_2)\} \right\} \right\}.$$

Given $x_2 \in M_2$, we introduce a new variable $e_1(x_2) \in \mathbb{R}$ such that $e_1(x_2) \leq f_1(x_1,x_2)$ for all $x_1 \in M_1$. Then we replace the constraint $\min_{x_i \in M_i, i=1,2,3} \{f(x_1,x_2,x_3)\} \geq 0$ by the constraint

$$\min_{x_3 \in M_3} \left\{ f_3(x_3) + \min_{x_2 \in M_2} \{f_2(x_2,x_3) + e_1(x_2)\} \right\} \geq 0, \tag{5.5a}$$

$$e_1(x_2) \leq f_1(x_1,x_2) \quad (x_1 \in M_1, x_2 \in M_2). \tag{5.5b}$$

Proposition 5.3.2. *If* $\min_{x_i \in M_i, i=1,2,3} \{f(x_1,x_2,x_3)\} \geq 0$*, then there is some* $e_1(x_2) \in \mathbb{R}$ *such that* $e_1(x_2) \leq f_1(x_1,x_2)$ *for all* $x_1 \in M_1$ *for every* $x_2 \in M_2$*. Conversely, if* (5.5) *holds, then* $\min_{x_i \in M_i, i=1,2,3} \{f(x_1,x_2,x_3)\} \geq 0$*.*

Proof. Let $x_2 \in M_2$. Then $e_1(x_2) \coloneqq \min_{x_1' \in M_1} f_1(x_1',x_2) \leq f_1(x_1,x_2)$ for all $x_1 \in M_1$. The second assertion follows immediately by substituting (5.5b) into (5.5a). □

By Proposition 5.3.2, the constraints $\min_{x_i \in M_i, i=1,2,3} \{f(x_1,x_2,x_3)\} \geq 0$ and (5.5) are equivalent. Iterating this procedure, we can rewrite constraint (5.5) by

$$\begin{aligned} &f_3(x_3) + e_2(x_3) \geq 0 \quad (x_3 \in M_3), \\ &e_1(x_2) \leq f_1(x_1,x_2) \quad (x_1 \in M_1, x_2 \in M_2), \\ &e_2(x_3) \leq f_2(x_2,x_3) + e_1(x_2) \quad (x_2 \in M_2, x_3 \in M_3). \end{aligned} \tag{5.6}$$

Note that constraint (5.6) is linear where the variables are $e_1(x_2)$, $x_2 \in M_2$, and $e_2(x_3)$, $x_3 \in M_3$. This procedure obviously increases the number of constraints. Constraint (5.6) consists of $|M_2| + |M_3|$ variables and $|M_3| + |M_1| \cdot |M_2| + |M_2| \cdot |M_3|$ linear constraints. Instead of evaluating f for every single $(x_1,x_2,x_3) \in M_1 \times M_2 \times M_3$, we can use the respective linear constraint to check its validity.

5.3.3. Numerical Example

In this section, we use the preceding approximation method to solve the accompanying airport example of chapter 3 where the staff size is two. Then we have

$$\tilde{\Sigma}(\mathrm{TS}) = \{\mathrm{TS}, \mathrm{T}\}, \quad \tilde{\Sigma}(\mathrm{T}) = \{\mathrm{TS}, \mathrm{T}, \mathrm{A}\}, \quad \tilde{\Sigma}(\mathrm{A}) = \{\mathrm{TS}, \mathrm{T}, \mathrm{A}, \mathrm{F}\}, \quad \tilde{\Sigma}(\mathrm{F}) = \{\mathrm{F}\}.$$

Here, we define basis functions which depend on one sector only. We consider basis functions for each sector which are supposed to be arbitrary polynomials of degree two. To this end, we define the constant function $h_0(s) \coloneqq 1$, $s \in S$, and for each sector σ the basis functions

$$h_{\sigma,1}(s) \coloneqq s(\sigma), \quad h_{\sigma,2}(s) \coloneqq s(\sigma)^2, \quad s \in S.$$

Let $(s,a) \in D$. Then we define for $\sigma \in \Sigma$

$$\begin{aligned} \mathscr{S}^a_\sigma(s) \coloneqq \sum_{j=1}^{2} w_{\sigma,j} \Bigg[&- \left(\alpha - \sum_{\sigma \in \tilde{\Sigma}(\sigma)} \lambda_\sigma(s(\sigma), a(\sigma)) \right) h_{\sigma,j}(s) \\ &+ \sum_{\sigma^* \in \tilde{\Sigma}(\sigma)} \left[\lambda_{a(\sigma^*)}(\sigma^*) \sum_{g' \in G} \Phi^{s(\sigma^*)}_{a(\sigma^*),\sigma^*}(g') h_{\sigma,j}\left(\zeta^{g'}_{a(\sigma^*),\sigma^*}(s) \right) + \sum_{e \in \mathscr{E}(\sigma^*)} \lambda_e(s(\sigma^*)) h_{\sigma,j}(\zeta_e(s)) \right] \Bigg], \end{aligned}$$

$s \in S$. Note that $\mathscr{S}^a_{\mathrm{TS}}(s) = \mathscr{S}^a_{\mathrm{TS}}(s')$ if $s(\mathrm{TS}) = s'(\mathrm{TS})$ and $s(\mathrm{T}) = s'(\mathrm{T})$, that $\mathscr{S}^a_{\mathrm{T}}(s) = \mathscr{S}^a_{\mathrm{T}}(s')$ if $s(\mathrm{TS}) = s'(\mathrm{TS})$, $s(\mathrm{T}) = s'(\mathrm{T})$ and $s(\mathrm{A}) = s'(\mathrm{A})$, and that $\mathscr{S}^a_{\mathrm{F}}(s) = \mathscr{S}^a_{\mathrm{F}}(s')$ if $s(\mathrm{F}) = s'(\mathrm{F})$ due to the assumptions on the $h_{\sigma,i}$, $\sigma \in \Sigma$, $i = 1,2$. Furthermore, $D(s) = D(s') =: \mathscr{D}$ for all s, $s' \in S$. Hence, for some fixed $\zeta : S \to \mathbb{R}_{>0}$ an approximate non-linear program is given by:

Maximize $\displaystyle\sum_{s\in S} \zeta(s) \left[w_0 + \sum_{\sigma\in\Sigma} \sum_{j=1}^{2} w_{\sigma,j} h_{\sigma,j}(s) \right]$

under the constraint

$$\min_{s(\mathrm{F})\in G} \left\{ \min_{s(\mathrm{A})\in G} \left\{ \min_{s(\mathrm{T})\in G} \left\{ \min_{s(\mathrm{TS})\in G} \{C_{\mathrm{TS}}(s(\mathrm{TS}),a(\mathrm{TS})) + \mathscr{S}^a_{\mathrm{TS}}(s) + \mathscr{S}^a_{\mathrm{T}}(s) + \mathscr{S}^a_{\mathrm{A}}(s)\} + C_{\mathrm{T}}(s(\mathrm{T}),a(\mathrm{T})) \right\} \right.\right.$$
$$\left.\left. + C_{\mathrm{A}}(s(\mathrm{A}),a(\mathrm{A})) \right\} + C_{\mathrm{F}}(s(\mathrm{F}),a(\mathrm{F})) + \mathscr{S}^a_{\mathrm{F}}(s) \right\} + w_0 \geq 0 \quad (a \in \mathscr{D}).$$

Assume that $\zeta \equiv 1/625$. Then we have for the objective function

$$\sum_{s\in S} \zeta(s) \left[w_0 + \sum_{\sigma\in\Sigma} \sum_{j=1}^{2} w_{\sigma,j} h_{\sigma,j}(s) \right] = w_0 + \frac{1}{625} \sum_{s\in S} \sum_{\sigma\in\Sigma} \sum_{j=1}^{2} w_{\sigma,j} h_{\sigma,j}(s)$$
$$= w_0 + \frac{1}{625} \sum_{\sigma\in\Sigma} w_{\sigma,1} \sum_{s\in S} h_{\sigma,1}(s) + \frac{1}{625} \sum_{\sigma\in\Sigma} w_{\sigma,2} \sum_{s\in S} h_{\sigma,2}(s) = w_0 + \sum_{\sigma\in\Sigma} 2 w_{\sigma,1} + \sum_{\sigma\in\Sigma} 6 w_{\sigma,2},$$

since $\sum_{s\in S} h_{\sigma,1}(s) = \sum_{s\in S} s(\sigma) = 125 \sum_{g\in G} g = 125 \cdot 10 = 1250$ and $\sum_{s\in S} h_{\sigma,2}(s) = 125 \sum_{g\in G} g^2 = 3750$, $\sigma \in \Sigma$. By the linearization technique of the preceding section, this is equivalent to the following approximate linear program:

Maximize $\displaystyle w_0 + \sum_{\sigma\in\Sigma} 2 w_{\sigma,1} + \sum_{\sigma\in\Sigma} 6 w_{\sigma,2}$

under the constraint

$$\begin{aligned}
&e^a_{\mathrm{TS}}(s(\mathrm{T}),s(\mathrm{A}),s(\mathrm{F})) \leq C_{\mathrm{TS}}(s(\mathrm{TS}),a(\mathrm{TS})) + \mathscr{S}^a_{\mathrm{TS}}(s) + \mathscr{S}^a_{\mathrm{T}}(s) + \mathscr{S}^a_{\mathrm{A}}(s) \quad (s(\mathrm{TS}),\, s(\mathrm{T}),\, s(\mathrm{A}),\, s(\mathrm{F}) \in G),\\
&e^a_{\mathrm{T}}(s(\mathrm{A}),s(\mathrm{F})) \leq C_{\mathrm{T}}(s(\mathrm{T}),a(\mathrm{T})) + e^a_{\mathrm{TS}}(s(\mathrm{T}),s(\mathrm{A}),s(\mathrm{F})) \quad (s(\mathrm{T}),\, s(\mathrm{A}),\, s(\mathrm{F}) \in G),\\
&e^a_{\mathrm{A}}(s(\mathrm{F})) \leq C_{\mathrm{A}}(s(\mathrm{A}),a(\mathrm{A})) + e^a_{\mathrm{T}}(s(\mathrm{A}),s(\mathrm{F})) \quad (s(\mathrm{A}),\, s(\mathrm{F}) \in G),\\
&e^a_{\mathrm{F}} \leq C_{\mathrm{F}}(s(\mathrm{F}),a(\mathrm{F})) + e^a_{\mathrm{A}}(s(\mathrm{F})) \quad (s(\mathrm{F}) \in G),\\
&e^a_{\mathrm{F}} + w_0 \geq 0\\
&\text{for all } a \in \mathscr{D}.
\end{aligned} \tag{5.7}$$

In (5.7), the variables are w_0, the $w_{\sigma,j}$ and the $e^a_{\mathrm{TS}}(s(\mathrm{T}),s(\mathrm{A}),s(\mathrm{F}))$, $e^a_{\mathrm{T}}(s(\mathrm{A}),s(\mathrm{F}))$, $e^a_{\mathrm{T}}(s(\mathrm{F}))$, e^a_4. A solution is given by

$$w^*_0 = 207{,}302.5, \quad w^*_{\mathrm{TS},1} = -919.5, \quad w^*_{\mathrm{TS},2} = 1{,}499.5, \quad w^*_{\mathrm{T},1} = -480.0, \quad w^*_{\mathrm{T},2} = 2{,}763.2, \quad w^*_{\mathrm{A},1} = -747.3,$$
$$w^*_{\mathrm{A},2} = 1{,}849.8, \quad w^*_{\mathrm{F},1} = -252.2, \quad w^*_{\mathrm{F},2} = 408.0,$$

so that $\hat{v}^2(0,0,0,0) = 207{,}302.5$, $\hat{v}^2(2,2,2,2) = 228{,}586.1$ and $\hat{v}^2(4,4,4,4) = 302{,}033.2$. We can see that the approximate value is quite far away from v^*. Using the greedy algorithm, we obtain an approximate policy $\hat{\mu}^2$ which leads to the value function $v^{\hat{\mu}^2}$ with

$$v^{\hat{\mu}^2}(0,0,0,0) = 38{,}550{,}016.2, \quad v^{\hat{\mu}^2}(2,2,2,2) = 40{,}418{,}014.8, \quad v^{\hat{\mu}^2}(4,4,4,4) = 68{,}077{,}128.3,$$

which is about 76 times worse than the optimal value function in $(0,0,0,0)$ and 25 times worse in $(4,4,4,4)$. The decision rule $\hat{\mu}^2$ given by the greedy algorithm with respect to $\hat{v}^2$ is illustrated in Figure 5.1. One can see that much more often elementary action 1 is chosen than in the optimal decision rule (cf. Figure 3.2).

A similar experiment was done by using polynomials of degree four as the basis functions. The value function of the greedy policy $\hat{\mu}^4$ takes the values

$$v^{\hat{\mu}^4}(0,0,0,0) = 2{,}143{,}461.8, \quad v^{\hat{\mu}^4}(2,2,2,2) = 2{,}563{,}478.2, \quad v^{\hat{\mu}^4}(4,4,4,4) = 21{,}061{,}804.4.$$

In $(0,0,0,0)$, the value function is about four times the optimal value, and in $(4,4,4,4)$, it is about eight times the optimal value. But the performance has obviously improved compared to the approximation with polynomials of degree two. The greedy policy with respect to $\hat{v}^4$ can be seen in Figure 5.2. Again, elementary action 1 is taken far more often than in an optimal decision rule.

Since the results of approximate linear programming are not very promising and since a one-step iteration might be hard to perform for the greedy algorithm, the concept of approximate linear programming has been discarded as a solution technique for the surveillance task.

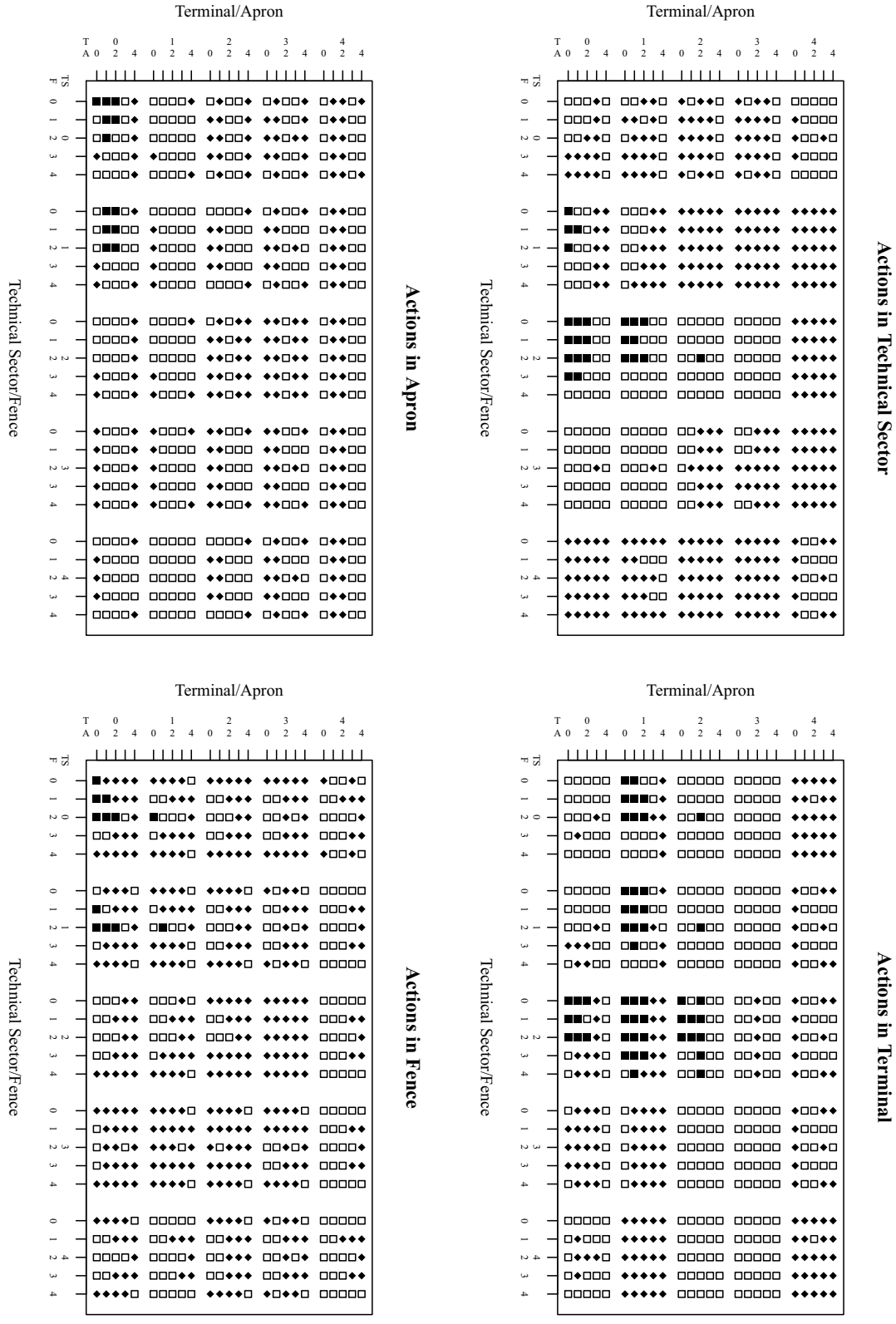


Figure 5.1.: Greedy policy with respect to $\hat{v}^2$.

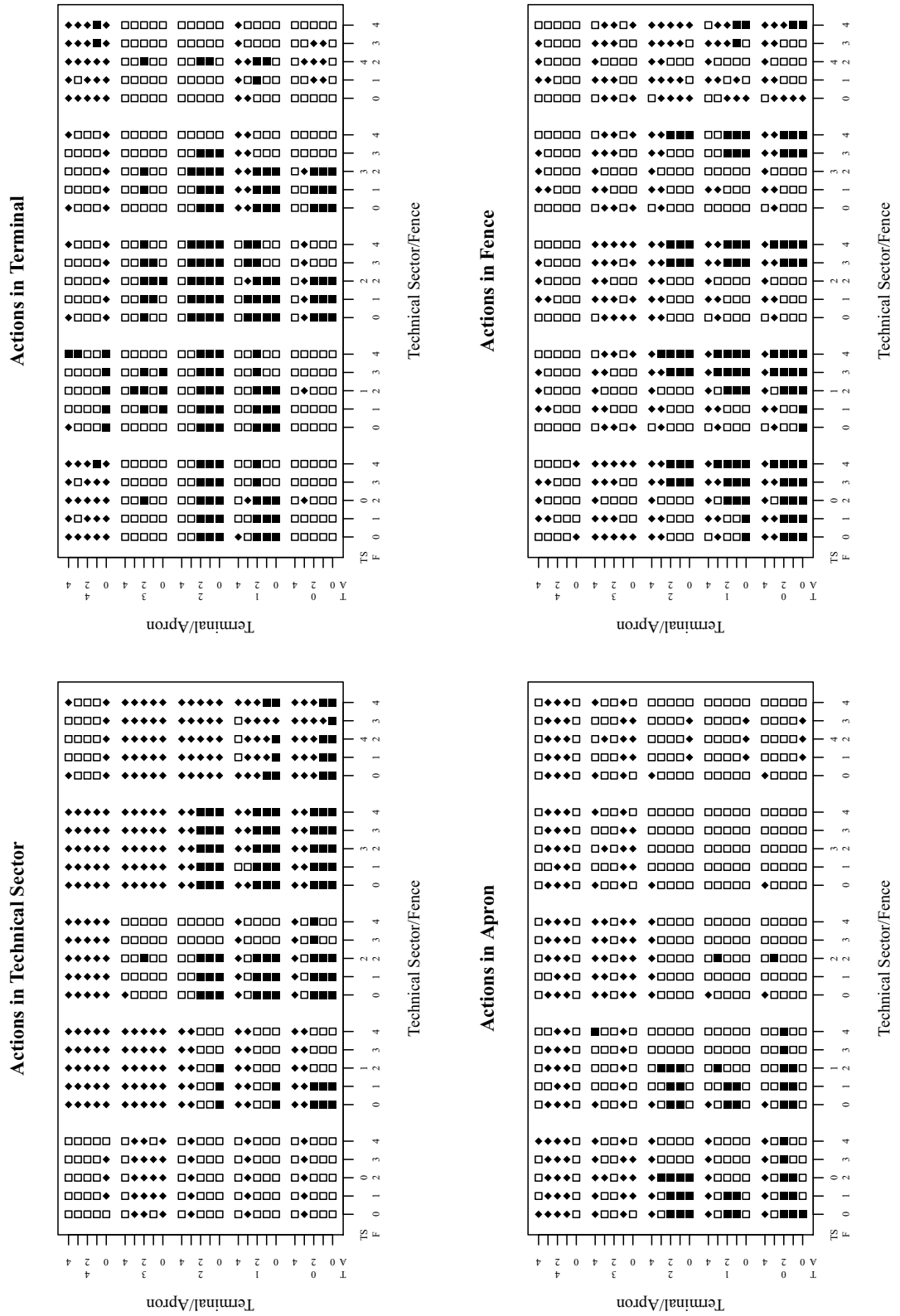


Figure 5.2.: Greedy policy with respect to $\hat{v}^4$.

5.4. Index-Based Heuristics

In this section, we present another approximate solution method for the surveillance task. The idea is quite simple: assume that the security staff consists of r persons. The question is: in which sectors should the security staff perform which elementary actions so that it is most beneficial for the whole infrastructure? This question will be answered by the definition of an appropriate index.

5.4.1. Index Rules

In this section, we present the concept of an index and the index rule based thereon. Assume that there are n projects (P_i), $i = 1, \ldots, n$, the decision maker has to work on. The dynamics of these projects are assumed to be independent from each other. In general, an index rule assigns a real number to each project depending on the state of the respective project only. This number is called the *index* of the respective project. A decision rule which assigns actions to those r projects having the highest indices is called the *index rule.*

The appeal of an index rule is its simplicity. One need not solve the system of all projects with the restriction that only r projects can be worked on. It is sufficient to derive the indices for each project solely, from which the index rule is easily determined. But the problem is to find appropriate indices. In general, indices giving satisfactory results need not exist for given projects.

In the following, we present two well-known indices for MDPs, the Gittins index and the Whittle index. Then we introduce a new heuristic index, which is easy to derive for the single projects. A small numerical study follows. Then we define the heuristics for the surveillance task. Finally, we examine the influence of the parameters of the surveillance task on its performance.

5.4.2. The Gittins Index

A well-known result from MDP-theory is the following, which was first presented in (Gittins, 1979). There is an abundance of further proofs available in the literature. The problem is originally formulated for MDPs in discrete time. For the definition of MDPs, we refer to section 7.2. Assume that we have n projects (P_i), $i = 1, \ldots, n$, which all are MDPs and which all are independent from each other. Furthermore, we assume that

- in each state, there are exactly two possible actions, a passive action p and an active action a,
- if the passive action p is chosen, the respective project remains in the current state and no cost arises,
- if the active action a is taken, the project transitions according to given transition probabilities and a cost only depending on the current state of the project arises,
- the decision maker can work on at most one project at the same time.

This model is called the *multi-armed bandit model.* For such a model, (Gittins, 1979) shows that there is an optimal decision rule of the following form: let $s_i \in S_i$ be the state of project (P_i), $i = 1, \ldots, n$. Then it is optimal with respect to the expected total discounted cost criterion to perform the active action a in project (P_i) with $\nu_i^G(s_i) = \max_{j=1,\ldots,n} \nu_j^G(s_j)$ if $\nu_i^G(s_i) \geq 0$, for appropriate functions $\nu_j^G : S_j \to \mathbb{R}$, which depend only on the parameters of (P_j), $j = 1, \ldots, n$. Otherwise, it is optimal to perform the passive action p in all projects. The function ν_j^G is called the *Gittins index with respect to project* (P_j).

5.4.3. The Whittle Index

The Whittle index is a generalization of the Gittins index. In (Whittle, 1988), a *restless bandit model* is considered. Again, we have n independent projects (P_i), all of which are assumed to be MDPs, with respective state spaces S_i, $i = 1, \ldots, n$. The projects behave like in the multi-armed bandit model only that the states might change under the passive action p as well and that there might be a cost for the passive action, too, i. e., for project (P_i), we have the cost function $C_i(s_i, a_i)$, $s_i \in S_i$, $a_i \in \{a, p\}$. In the restless bandit model, it is assumed that the decision maker is forced to work on exactly r projects at the same time. (Whittle, 1988) derives an index by relaxing the requirement of working on exactly r projects at the same time by the constraint that in average on r projects are worked on simultaneously when considering the expected average cost criterion. For the expected total discounted cost criterion, the appropriate relaxation is that policies π are considered such that in discounted average $E^\pi\left[\sum_{k=0}^{\infty} \beta^k R_k\right] = r/(1-\beta)$ projects have to be worked on, where $\beta \in (0,1)$

is the respective discount factor and R_k is the number of projects being worked on at time step k (cf. section 7.2). So, for a given initial state $x_0 = (x_{1,0}, \ldots, x_{n,0}) \in \bigtimes_{i=1}^{n} S_i$, one has the Lagrangian function L for the relaxed problem given by

$$L(\pi, \nu) := \sum_{i=1}^{n} E^{\pi}\left[\sum_{k=0}^{\infty} \beta^k \left(C_i(X_{i,k}, A_{i,k}) + \nu R_{i,k}\right) \,\middle|\, X_{i,0} = x_{i,0}\right] - \frac{\nu r}{1-\beta}, \quad \pi \in \Pi,\ \nu \in \mathbb{R}, \tag{5.8}$$

where Π is the set of all randomized history-dependent policies for MDPs of the restless bandit model, $(X_k)_{k\in\mathbb{N}_0} = ((X_{1,k}, \ldots, X_{n,k}))_{k\in\mathbb{N}_0}$ is the state process of the restless bandit and $(R_{i,k})_{k\in\mathbb{N}_0}$ is the process which indicates that the active action a is taken for project (P_i), $i = 1, \ldots, n$, with tax ν on the active action, i. e., $R_{i,k} = 1$ if $A_{i,k} = a$ and $R_{i,k} = 0$ if $A_{i,k} = p$, $k \in \mathbb{N}_0$. The Langrangian multiplier ν can be seen as a tax on the active action a for each project since it is only added if the respective project is active. For fixed $\nu \in \mathbb{R}$, the underlying projects (P_i) are now decoupled. Minimizing $L(\cdot, \nu)$ can be solved by considering the underlying MDPs (P_i) with the tax ν on the active action solely. From MDP-theory, it is known that a decision rule $\mu_i^{\nu,*}$ is optimal for (P_i), $i = 1, \ldots, n$, cf. Theorem 6.2.10 of (Puterman, 2005). Now, we come to the definition of the Whittle index. But at first, we need the notion of indexability.

Definition 5.4.1. A project is *indexable* if the set of states in which the active action a is optimal is increasing from the empty set to the full state space when letting ν from ∞ to $-\infty$.

General projects need not be indexable as (Whittle, 1988) demonstrates by an example. For a state s_i of an indexable project (P_i), the Whittle index is defined as the unique tax $\nu_i^W(s_i)$ such that the passive action p and the active action a are equally preferable to the decision maker. Note that the tax on the active action a is raised at all states of the project (P_i). If $\nu^* \neq 0$ minimizes (5.8), then it is optimal with respect to the relaxed problem to act in every project (P_i) according to a deterministic stationary policy $\mu_i^{\nu^*,*}$ where the tax ν^* is raised for every (P_i), $i = 1, \ldots, n$. If all projects are indexable, then it is optimal with respect to the relaxed criterion to use the active action a in project (P_i) if the current state $s_i \in S_i$ satisfies $\nu_i^W(s_i) > \nu^*$ and to use the passive action in (P_i) if the current state $s_i \in S_i$ satisfies $\nu_i^W(s_i) < \nu^*$.

Again, one can define a policy based on the index in order to approximate an optimal decision rule for the problem with the strict constraint on working on exactly r projects: perform the active action a on those r projects having the largest indices, and perform the passive action p on all other projects. This policy need not be optimal in general. But computational experiments have shown that this policy is a very good heuristics (cf., e. g., (Ansell et al., 2003)). Moreover, (Whittle, 1988) shows that the Whittle index is equivalent to the Gittins index if the restless bandit model is degenerated to the multi-armed bandit model.

The Whittle index can easily be carried over to CMDPs by assigning a tax in form of an additional cost rate on the active action a. The relaxed constraint is then $E^{\pi}\left[\int_0^{\infty} e^{-\alpha t} R_t \, dt \mid X_0 = x_0\right] = r/\alpha$, $\pi \in \Pi$, where $\alpha > 0$ is the discount rate and $(R_t)_{t\geq 0}$ is the process of the number of allocated resources or the number of projects which are currently worked on.

A disadvantage of the Whittle index is that it may not be easily derived for any indexable project. For some specific models, the Whittle index is known (cf. (Whittle, 1988; Ansell et al., 2003)). For a certain CMDP model, the Whittle index is derived in (Dusonchet and Hongler, 2003). Especially for the surveillance task, the Whittle index is hard to derive due to the complexity of the model, assuming there are only two elementary actions for each sector.

5.4.4. A Heuristic Index

Although the Whittle index is hard to compute in general, the idea of assigning an index or some kind of measure to the projects remains because a policy based on such an index is easy to implement. For the surveillance task, the idea is the following: we split the infrastructure into its subinfrastructures of a given size, from which the respective surveillance task can be exactly computed in an adequate time. Then we assign the resources to such a subinfrastructure that benefits the most from allocating the security staff.

In this section, we define an index for a specific CMDP-model which includes the surveillance task of chapter 3. Furthermore, we show that this index has meaningful interpretations. Throughout this section, we assume that the individual projects are of the form as given in Assumption 5.4.2.

Assumption 5.4.2. Let $\Gamma = (S, A, D, \Lambda, P, \mathscr{K}, \mathscr{C}, \alpha)$ be a CMDP with the following properties:

- Let $K(s,a) = 0$ for all $(s,a) \in D$.
- There is a passive action $p \in D(s)$ for every $s \in S$.
- For the cost rates, we have $C(s,a) = C^{\text{state}}(s) + C^{\text{action}}(a)$ for all $(s,a) \in D$ and $C^{\text{action}}(p) = 0$.
- For the transition rates, we have $\lambda(s,a) = \lambda^{\text{state}}(s) + \lambda^{\text{action}}(a)$ for all $(s,a) \in D$ and $\lambda^{\text{action}}(p) = 0$.

- There are transition probabilities $p_{ss'}$ driven only by the current state $s \in S$, and there are transition probabilities $p^a_{ss'}$ driven by the current active action $a \in D(s)$ such that the Bellman equation for the optimal value function v^* can be written as

$$v(s) = \min_{a \in D(s)} \left\{ \frac{C^{\text{state}}(s) + C^{\text{action}}(a) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v(s') + \lambda^{\text{action}}(a) \sum_{s' \in S} p^a_{ss'} v(s')}{\lambda^{\text{state}}(s) + \lambda^{\text{action}}(a) + \alpha} \right\}, \quad s \in S,$$

  where $v : S \to \mathbb{R}$.

Note that we allow numerous active actions in Assumption 5.4.2. The costs are given by the penalty cost rate for being in state s by $C^{\text{state}}(s)$ and the cost rate $C^{\text{action}}(a)$ for executing action a. A project satisfying Assumption 5.4.2 has an inner dynamics driven by the current state which cannot be turned off. But this dynamics can be modified by the current action. This means, whatever action the decision maker chooses, the system has the same drive to reach subsequent states as under the passive action p. This is just the same as in the surveillance task of chapter 3. Assume that in a certain sector, the threat level is g. Then threat events occur independently of the current action at given rates. The decision maker can only hope that the current action is accomplished before a threat event occurs, but active actions cannot modify the entrance rates of the threat events.

If $D(s) = \{a, p\}$ for all $s \in S$, the model of Assumption 5.4.2 is a specific restless bandit model. But the model of Assumption 5.4.2 generalizes a restless bandit model in allowing to model more than just one active action.

Furthermore, note that any infrastructure modelled in terms of chapter 3 satisfies Assumption 5.4.2. One only has to identify the passive action p with the passive elementary action 0 of the surveillance task.

At first, we have a look at single projects satisfying Assumption 5.4.2.

Lemma 5.4.3. *Let (P) be a project satisfying Assumption 5.4.2. Let v^* be the optimal value function of (P), and let μ^* be an optimal decision rule for (P). If p is optimal in $s \in S$, then it follows*

$$\lambda^{\text{action}}(a) \left(v^*(s) - \sum_{s' \in S} p^a_{ss'} v^*(s') \right) - C^{\text{action}}(a) \le 0 \quad (a \in D(s)). \tag{5.9}$$

Furthermore, we have in general

$$\iota(s) := \lambda^{\text{action}}(\mu^*(s)) \left(T_p v^*(s) - \sum_{s' \in S} p^{\mu^*(s)}_{ss'} v^*(s') \right) - C^{\text{action}}(\mu^*(s)) \ge 0, \quad s \in S. \tag{5.10}$$

Proof. At first, we define for $v : S \to \mathbb{R}$ the usual one-step cost operators for CMDPs by

$$T_a v(s) := \frac{C^{\text{state}}(s) + C^{\text{action}}(a) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v(s') + \lambda^{\text{action}}(a) \sum_{s' \in S} p^a_{ss'} v(s')}{\lambda^{\text{state}}(s) + \lambda^{\text{action}}(a) + \alpha}, \quad a \in D(s), \quad \text{and}$$

$$T v(s) := \min_{a \in D(s)} T_a v(s), \quad s \in S.$$

By Theorem 2.3.4, p is optimal in $s \in S$ if and only if

$$\begin{aligned}
& T_p v^*(s) = T v^*(s) \quad \Leftrightarrow \quad T_p v^*(s) \le T_a v^*(s) \quad (a \in D(s)) \\
\Leftrightarrow \quad & \frac{C^{\text{state}}(s) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v^*(s')}{\lambda^{\text{state}}(s) + \alpha} \\
& \qquad \le \frac{C^{\text{state}}(s) + C^{\text{action}}(a) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v^*(s') + \lambda^{\text{action}}(a) \sum_{s' \in S} p^a_{ss'} v^*(s')}{\lambda^{\text{state}}(s) + \lambda^{\text{action}}(a) + \alpha} \quad (a \in D(s)) \\
\Leftrightarrow \quad & \lambda^{\text{action}}(a) \left(C^{\text{state}}(s) + \lambda^{\text{state}}(s) \sum_{s \in S} p_{ss'} v^*(s') \right) \\
& \qquad \le \left(\lambda^{\text{state}}(s) + \alpha \right) \left(C^{\text{action}}(a) + \lambda^{\text{action}}(a) \sum_{s' \in S} p^a_{ss'} v^*(s') \right) \quad (a \in D(s)) \\
\Leftrightarrow \quad & \lambda^{\text{action}}(a) \Bigg(\underbrace{\frac{C^{\text{state}}(s) + \lambda^{\text{state}}(s) \sum_{s \in S} p_{ss'} v^*(s')}{\lambda^{\text{state}}(s) + \alpha}}_{= T_p v^*(s)} - \sum_{s' \in S} p^a_{ss'} v^*(s') \Bigg) - C^{\text{action}}(a) \le 0 \quad (a \in D(s)).
\end{aligned} \tag{5.11}$$

Inequality (5.9) follows from the assumption that p is optimal in s, and therefore $T_p v^*(s) = Tv^*(s) = v^*(s)$. Inequality (5.10) follows from $T_{\mu^*(s)} v^*(s) \leq T_p v^*(s)$, $s \in S$, and from similar manipulations yielding (5.11) with opposite inequality sign. □

Let us consider $\iota(s)$, which is defined in Lemma 5.4.3, for some $s \in S$. Assume that $\lambda(\mu^*(s)) > 0$ and $C^{\text{action}}(\mu^*(s)) \geq 0$. Then, we have $T_p v^*(s) - \sum_{s' \in S} p_{ss'}^{\mu^*(s)} v^*(s) \geq 0$. Heuristically, increasing $\lambda^{\text{action}}(\mu^*(s))$ or decreasing $C^{\text{action}}(\mu^*(s))$ both lead to an increase of $\iota(s)$ if all other parameters are the same. Of course, this is not mathematically correct since by changing $\lambda^*(\mu^*(s))$ or $C^{\text{action}}(\mu^*(s))$ also v^* changes. However, assuming we have two projects and we can only work on one of them, we would prefer to work on the one which transitions faster or which is cheaper if both have similar costs. So, the number $\iota(s)$ gives some information about the benefit from executing $\mu^*(s)$ compared to the passive action p in a project being in state $s \in S$. Therefore, we use ι as a heuristic index.

We can rewrite ι in the following way which gives more insight into how ι works:

Proposition 5.4.4. *Let (P) be a project satisfying Assumption 5.4.2, then we have*

$$\iota(s) = (\lambda(s, \mu^*(s)) + \alpha)\,(T_p v^*(s) - v^*(s)), \quad s \in S,$$

where v^* is the optimal value function of (P), and μ^* is an optimal decision rule for (P).

Proof. Since $v^* = Tv^*$ and μ^* is optimal, we compute for $s \in S$

$$\begin{aligned}
&v^*(s) + \frac{\iota(s)}{\lambda(s, \mu^*(s)) + \alpha} \\
&\quad = \frac{C^{\text{state}}(s) + C^{\text{action}}(\mu^*(s)) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v^*(s') + \lambda^{\text{action}}(\mu^*(s)) \sum_{s' \in S} p_{ss'}^{\mu^*(s)} v^*(s')}{\lambda(s, \mu^*(s)) + \alpha} \\
&\qquad + \frac{\lambda^{\text{action}}(\mu^*(s)) \left(T_p v^*(s) - \sum_{s' \in S} p_{ss'}^{\mu^*(s)} v^*(s) \right) - C^{\text{action}}(\mu^*(s))}{\lambda(s, \mu^*(s)) + \alpha} \\
&\quad = \frac{C^{\text{state}}(s) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v^*(s') + \lambda^{\text{action}}(\mu^*(s))\, T_p v^*(s)}{\lambda(s, \mu^*(s)) + \alpha} \\
&\quad = \frac{(\lambda^{\text{state}}(s) + \alpha)\, T_p v^*(s) + \lambda^{\text{action}}(\mu^*(s))\, T_p v^*(s)}{\lambda(s, \mu^*(s)) + \alpha} = T_p v^*(s),
\end{aligned}$$

from which the assertion follows. □

Remark 5.4.5. The factor $(\lambda(s, \mu^*(s)) + \alpha)$ approximately gives the rate until the next state transition when using the optimal action $\mu^*(s)$ in $s \in S$ if α is near zero. So, the larger this factor, the faster a resource working on the respective project would be deallocated.

The second factor $(T_p v^*(s) - v^*(s))$ compares two policies: v^* is the value function of an optimal policy. Whereas $T_p v^*(s)$ can be seen as an approximation of a policy which acts optimally in all states except for s, in which the passive action p is chosen. Motivated by the value iteration method of Proposition 2.3.6, it is a common approach in approximate dynamic programming to obtain an approximation of the value function for some given decision rule $\mu : S \to A$ by starting with some initial approximating value function $v_0 : S \to \mathbb{R}$. Then a one-step iteration is applied to v_0 so that we have a new approximation v_1 given by

$$v_1(s) = \frac{C(s, \mu(s)) + \lambda(s, \mu(s)) \sum_{s' \in S} p_{ss'}^{\mu(s)} v_0(s')}{\lambda(s, \mu(s)) + \alpha}, \quad s \in S.$$

This is essentially the same as the greedy algorithm with respect to v_0 except that here the restriction set is degenerated to $D(s) = \{\mu(s)\}$, $s \in S$.

Together, the index $\iota(s)$ measures the approximate benefit from using the optimal action instead of the passive action p in state s.

Another argument for using ι as an index is due to the following theorem.

Theorem 5.4.6. *Let (P) be a project satisfying Assumption 5.4.2. Then we have for $s \in S$*

1. *$\iota(s) \geq 0$, and*
2. *$\iota(s) = 0$ if and only if p is optimal in s.*

Proof. The first assertion is proved in Lemma 5.4.3. Since $\alpha > 0$, the second assertion follows from $v^* = Tv^* \leq T_p v^*$ and $v^*(s) = T_p v^*(s)$ if and only if p is optimal in $s \in S$. □

By this theorem, when using ι as an index, it has the advantage of detecting projects in which it is optimal to perform the passive action. Hence, no resource would be wasted by allocating it to a project in which it is optimal to do nothing.

Remark 5.4.7. Note that any index $\hat{\iota}$ of the form $\hat{\iota}(s) = f(s)\,(T_p v^*(s) - v^*(s))$, $s \in S$, for some function $f : S \to \mathbb{R}_{>0}$, satisfies the assertions of Theorem 5.4.6. Especially, we define the following modified heuristic index

$$\tilde{\iota}(s) \coloneqq T_p v^*(s) - v^*(s), \quad s \in S,$$

which only measures the benefit from using the optimal action without taking the transition rate under the optimal action $\lambda(s, \mu^*(s))$ into consideration explicitly. Of course, the transition rate is implicitly used in order to determine v^*.

From the proof of Proposition 5.4.4, we obtain the next corollary.

Corollary 5.4.8. *Let (P) be a project satisfying Assumption 5.4.2. Let $(\tilde{P})$ be a project of the following form: the restriction set is $\tilde{D}(s) \coloneqq \{0, \mu^*(s)\}$, $s \in S$, where μ^* is an optimal decision rule for (P), and all further parameters are the same except that the costs are given by $\tilde{C}(s, \mu^*(s)) = C(s, \mu^*(s)) + \iota(s)$ for all $s \in S$. Then*

$$\tilde{T}_{\mu^*(s)} v^*(s) = \tilde{T}_p v^*(s) \quad (s \in S),$$

where v^* is the optimal value function of (P) and $\tilde{T}_p$ and $\tilde{T}_{\mu^*(s)}$ are the one-step cost operators for $(\tilde{P})$ for action p and $\mu^*(s)$, $s \in S$, respectively, i. e.,

$$\tilde{T}_p v(s) \coloneqq \frac{C^{\text{state}}(s) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'}\, v(s')}{\lambda^{\text{state}}(s) + \alpha},$$

$$\tilde{T}_{\mu^*(s)} v(s) \coloneqq \frac{C^{\text{state}}(s) + C^{\text{action}}(\mu^*(s)) + \iota(s) + \lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'}\, v(s') + \lambda^{\text{action}}(\mu^*(s)) \sum_{s' \in S} p_{ss'}^{\mu^*(s)}\, v(s')}{\lambda^{\text{state}}(s) + \lambda^{\text{action}}(\mu^*(s)) + \alpha}, \quad s \in S,$$

for $v : S \to \mathbb{R}$.

Proof. For $s \in S$, we note that $\iota(s) = 0$ if and only if $\mu^*(s) = p$. Thus, the operators $\tilde{T}_p$ and $\tilde{T}_{\mu^*(s)}$, $s \in S$, are well-defined. Since $\tilde{T}_p = T_p$, the rest of the proof is the same as for Proposition 5.4.4. □

Remark 5.4.9. Corollary 5.4.8 gives another interpretation of the index. Assume that the optimal value functions of (P) and $(\tilde{P})$ are approximately the same. Then the decision maker is indifferent between the passive and the optimal action $\mu^*(s)$ in $(\tilde{P})$. Or in other words, $\iota(s)$ is the amount that when added to the cost rate of the optimal action makes the passive and the optimal action equally preferable to the decision maker. This can be seen as an approximate local version of the Whittle index. It is just an approximation since v^* is not the exact value function of $(\tilde{P})$, but it is considered to be an approximation of it, and ι is local since there is only a tax depending on the optimal action $\mu^*(s)$ at state s which might not be the same for all states $s \in S$.

If ι is defined as in Proposition 5.4.4, ι could be used as an index for more general types of projects. Then the interpretation of ι is just the same as in Remark 5.4.5: it is the approximate benefit from using the optimal action instead of the passive action. But ι cannot be interpreted as an approximation of the Whittle index as outlined in Remark 5.4.9.

5.4.5. Whittle Index Versus Heuristic Index

In this section, we present a small numerical study in which we compare the Whittle index with the indices ι and $\tilde{\iota}$ defined in the preceding section. To this end, we consider a specific model which could represent a deteriorating machine which can be repaired. It has two states, a good state and a bad one, and the decision maker has to decide whether she wants to repair the machine and if so, which of the machines she wants to be repaired.

Assumption 5.4.10. Assume that we have a project (P) of the following form: the project has two states 0 and 1. In each state, the decision maker can perform the passive action p and the active action a. If the project remains inactive, it transitions from 0 to 1 at a rate of $\lambda(0, p) = \lambda^{\text{state}}(0) =: \lambda_p > 0$, and from 1 to 0 at a rate of 0, i. e., $\lambda(1, p) = \lambda^{\text{state}}(1) = 0$, so that it remains in 1 in this case. If the active action is used, the project transitions from 0 to 0 at a rate of $\lambda_a > 0$ where λ_a is the rate of the active action and from 0 to 1 at rate λ_p so that $\lambda(0, a) = \lambda^{\text{state}}(0) + \lambda^{\text{action}}(a) = \lambda_p + \lambda_a$. (In order to satisfy Assumption 5.4.2, the artificial transition from 0 to 0 has to be defined.) Under the active action, the project transitions from 1 to 0 at a rate of $\lambda(1, a) = \lambda^{\text{state}}(1) + \lambda^{\text{action}}(a) = \lambda_a$. We assume that the cost rate of the active action for project (P) is $C^{\text{action}}(a) \geq 0$. There is no cost for being in state 0, i. e., $C^{\text{state}}(0) = 0$, and a cost rate $C^{\text{state}}(1) \geq 0$ for being in state 1. The dynamics of the model can be seen from Figure 5.3.

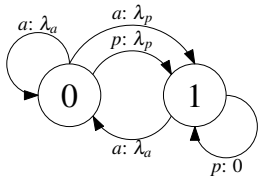


Figure 5.3.: Model of a project satisfying Assumption 5.4.10.

Note that a project satisfying Assumption 5.4.10 also satisfies Assumption 5.4.2. We might think of the project as of a production machine which can be in two states. It is either in the "good" state 0 or in the "bad" state 1 with which a certain cost rate $C^{\text{state}}(1)$ is associated, e. g., for production losses. It transitions from 0 to 1 at a given rate λ_p. In state 1, the decision maker can arrange a repair at a cost rate $C^{\text{action}}(a)$, or she could use the passive action leaving the machine in state 1. The rate λ_p can be seen as a deterioration rate of this machine and λ_a as the repair rate. If the decision maker has several such machines with different parameters, she has to decide which machines should be repaired now and which should be repaired at a later time and which should never be repaired. For such a project, we are able to derive the Whittle index explicitly.

Proposition 5.4.11. *A project (P) satisfying Assumption 5.4.10 is indexable, and the Whittle index ν^W takes the form*

$$\nu^W(0) = -C^{\text{action}}(a) \quad \textit{and} \quad \nu^W(1) = \frac{\lambda_a}{\lambda_p+\alpha} C^{\text{state}}(1) - C^{\text{action}}(a).$$

Proof. Recall that the Whittle index for state s is the tax on the active action a such that the passive and the active action are equally attractive to the decision maker. The Bellman equation for the optimal value function v_ν^* for the project with a tax $\nu \in \mathbb{R}$ on the active action is given by

$$v_\nu^*(0) = \min\left\{ \underbrace{\frac{1}{\lambda_p+\alpha}\lambda_p v_\nu^*(1)}_{\text{action } p}, \underbrace{\frac{1}{\lambda_p+\lambda_a+\alpha}\left(C^{\text{action}}(a)+\nu+\lambda_p v_\nu^*(1)+\lambda_a v_\nu^*(0)\right)}_{\text{action } a} \right\}, \tag{5.12a}$$

$$v_\nu^*(1) = \min\left\{ \underbrace{\frac{1}{\alpha}C^{\text{state}}(1)}_{\text{action } p}, \underbrace{\frac{1}{\lambda_a+\alpha}\left(C^{\text{state}}(1)+C^{\text{action}}(a)+\nu+\lambda_a v_\nu^*(0)\right)}_{\text{action } a} \right\}. \tag{5.12b}$$

From the Bellman equation (5.12), we obtain that p is optimal in state 0 if and only if

$$\begin{aligned}&(\lambda_p+\lambda_a+\alpha)\lambda_p v_\nu^*(1) \le (\lambda_p+\alpha)\left(C^{\text{action}}(a)+\nu+\lambda_p v_\nu^*(1)+\lambda_a v_\nu^*(0)\right)\\ &\quad\Leftrightarrow\quad \nu \ge \frac{\lambda_p\lambda_a}{\lambda_p+\alpha} v_\nu^*(1) - \lambda_a v_\nu^*(0) - C^{\text{action}}(a).\end{aligned} \tag{5.13}$$

Further, p is optimal in state 1 if and only if

$$\nu \ge -\lambda_a v_\nu^*(0) + \frac{\lambda_a}{\alpha} C^{\text{state}}(1) - C^{\text{action}}(a). \tag{5.14}$$

In the next steps, we compute the taxes ν for which the respective four deterministic stationary policies are optimal:

1. At first, we obtain the taxes ν for which it is optimal to use the decision rule μ^{pp}, defined by $\mu^{pp}(0) = \mu^{pp}(1) = p$. The value function of μ^{pp} with tax ν is given by

   $$v_\nu^{\mu^{pp}}(1) = \frac{1}{\alpha}C^{\text{state}}(1) \quad \text{and} \quad v_\nu^{\mu^{pp}}(0) = \frac{\lambda_p}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1).$$

   Then μ^{pp} is optimal if and only if

   $$\begin{aligned}\nu &\overset{(5.13)}{\ge} \frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) - \frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) - C^{\text{action}}(a) = -C^{\text{action}}(a) \quad \text{and}\\ \nu &\overset{(5.14)}{\ge} -\lambda_a \frac{\lambda_p}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) + \frac{\lambda_a}{\alpha} C^{\text{state}}(1) - C^{\text{action}}(a) = \frac{\lambda_a}{\lambda_p+\alpha} C^{\text{state}}(1) - C^{\text{action}}(a),\end{aligned}$$

   which holds if and only if $\nu \ge \max\{-C^{\text{action}}(a), \lambda_a/(\lambda_p+\alpha) C^{\text{state}}(1) - C^{\text{action}}(a)\}$.

2. Next, we consider the decision rule μ^{ap}, defined by $\mu^{ap}(0) = a$ and $\mu^{ap}(1) = p$. The associated value function is

$$v_\nu^{\mu^{ap}}(1) = \frac{1}{\alpha} C^{\text{state}}(1) \quad \text{and} \quad v_\nu^{\mu^{ap}}(0) = \frac{\lambda_p}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) + \frac{1}{\lambda_p+\alpha}\left(C^{\text{action}}(a)+\nu\right).$$

Hence, μ^{ap} is optimal if and only if

$$\begin{aligned}
\nu &\overset{(5.13)}{\leq} \frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) - \frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) - \frac{\lambda_a}{\lambda_p+\alpha}\left(C^{\text{action}}(a)+\nu\right) - C^{\text{action}}(a) \\
&\Leftrightarrow \quad \nu \leq -C^{\text{action}}(a) \quad \text{and} \\
\nu &\overset{(5.14)}{\geq} -\frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\alpha)} C^{\text{state}}(1) - \frac{\lambda_a}{\lambda_p+\alpha}\left(C^{\text{action}}(a)+\nu\right) + \frac{\lambda_a}{\alpha} C^{\text{state}}(1) - C^{\text{action}}(a) \\
&\Leftrightarrow \quad \nu \geq \frac{\lambda_a}{\lambda_p+\lambda_a+\alpha} C^{\text{state}}(1) - C^{\text{action}}(a),
\end{aligned}$$

which holds if and only if $\lambda_a/(\lambda_p+\lambda_a+\alpha) C^{\text{state}}(1) - C^{\text{action}}(a) \leq \nu \leq -C^{\text{action}}(a)$.

3. For the decision rule μ^{pa}, defined by $\mu^{pa}(0) = p$ and $\mu^{pa}(1) = a$, we obtain from (5.12a)

$$v_\nu^{\mu^{pa}}(0) = \frac{\lambda_p}{\lambda_p+\alpha} v_\nu^{\mu^{pa}}(1). \tag{5.15}$$

Further, (5.12b) yields

$$\begin{aligned}
&(\lambda_a+\alpha) v_\nu^{\mu^{pa}}(1) = C^{\text{state}}(1) + C^{\text{action}}(a) + \nu + \frac{\lambda_p\lambda_a}{\lambda_p+\alpha} v_\nu^{\mu^{pa}}(1) \\
&\Leftrightarrow \quad v_\nu^{\mu^{pa}}(1) = \frac{\lambda_p+\alpha}{\alpha(\lambda_p+\lambda_a+\alpha)} C^{\text{state}}(1) + \frac{\lambda_p+\alpha}{\alpha(\lambda_p+\lambda_a+\alpha)}\left(C^{\text{action}}(a)+\nu\right).
\end{aligned} \tag{5.16}$$

Substituting (5.16) into (5.15), yields

$$v_\nu^{\mu^{pa}}(0) = \frac{\lambda_p}{\alpha(\lambda_p+\lambda_a+\alpha)} C^{\text{state}}(1) + \frac{\lambda_p}{\alpha(\lambda_p+\lambda_a+\alpha)}\left(C^{\text{action}}(a)+\nu\right).$$

Therefore, the policy μ^{pa} is optimal if and only if

$$\begin{aligned}
\nu &\overset{(5.13)}{\geq} \frac{\lambda_p\lambda_a}{\lambda_p+\alpha} v_\nu^{\mu^{pa}}(1) - \frac{\lambda_p\lambda_a}{\lambda_p+\alpha} v_\nu^{\mu^{pa}}(1) - C^{\text{action}}(a) = -C^{\text{action}}(a) \quad \text{and} \\
\nu &\overset{(5.14)}{\leq} -\frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\lambda_a+\alpha)} C^{\text{state}}(1) - \frac{\lambda_p\lambda_a}{\alpha(\lambda_p+\lambda_a+\alpha)}\left(C^{\text{action}}(a)+\nu\right) + \frac{\lambda_a}{\alpha} C^{\text{state}}(1) - C^{\text{action}}(a) \\
&\Leftrightarrow \quad (\lambda_p+\alpha)(\lambda_a+\alpha)\nu \leq \lambda_a(\lambda_a+\alpha) C^{\text{state}}(1) - (\lambda_p+\alpha)(\lambda_a+\alpha) C^{\text{action}}(a) \\
&\Leftrightarrow \quad \nu \leq \frac{\lambda_a}{\lambda_p+\alpha} C^{\text{state}}(1) - C^{\text{action}}(a),
\end{aligned}$$

which holds if and only if $-C^{\text{action}}(a) \leq \nu \leq \lambda_a/(\lambda_p+\alpha) C^{\text{state}}(1) - C^{\text{action}}(a)$.

4. Last, for the decision μ^{aa}, defined by $\mu^{aa}(0) = \mu^{aa}(1) = a$, we obtain from (5.12a)

$$v_\nu^{\mu^{aa}}(0) = \frac{\lambda_p}{\lambda_p+\alpha} v_\nu^{\mu^{aa}}(1) + \frac{1}{\lambda_p+\alpha}\left(C^{\text{action}}(a)+\nu\right), \tag{5.17}$$

and from (5.12b), we obtain

$$\begin{aligned}
&(\lambda_a+\alpha) v_\nu^{\mu^{aa}}(1) = C^{\text{state}}(1) + C^{\text{action}}(a) + \nu + \lambda_a v_\nu^{\mu^{aa}}(0) \\
&\quad = C^{\text{state}}(1) + C^{\text{action}}(a) + \nu + \frac{\lambda_p\lambda_a}{\lambda_p+\alpha} v_\nu^{\mu^{aa}}(1) + \frac{\lambda_a}{\lambda_p+\alpha}\left(C^{\text{action}}(a)+\nu\right) \\
&\Leftrightarrow \quad (\lambda_p+\alpha)(\lambda_a+\alpha) v_\nu^{\mu^{aa}}(1) - \lambda_p\lambda_a v_\nu^{\mu^{aa}}(1) = (\lambda_p+\alpha) C^{\text{state}}(1) + (\lambda_p+\lambda_a+\alpha)\left(C^{\text{action}}(a)+\nu\right)
\end{aligned}$$

$$\Leftrightarrow \quad v_\nu^{\mu^{aa}}(1) = \frac{\lambda_p + \alpha}{\alpha(\lambda_p + \lambda_a + \alpha)} C^{\text{state}}(1) + \frac{1}{\alpha}\left(C^{\text{action}}(a) + \nu\right). \tag{5.18}$$

Substituting (5.18) into (5.17), yields

$$v_\nu^{\mu^{aa}}(0) = \frac{\lambda_p}{\alpha(\lambda_p + \lambda_a + \alpha)} C^{\text{state}}(1) + \frac{1}{\alpha}\left(C^{\text{action}}(a) + \nu\right).$$

Then μ^{aa} is optimal if and only if

$$\begin{aligned}
\nu &\overset{(5.13)}{\leq} \frac{\lambda_p \lambda_a}{\alpha(\lambda_p + \lambda_a + \alpha)} C^{\text{state}}(1) + \frac{\lambda_p \lambda_a}{\alpha(\lambda_p + \alpha)}\left(C^{\text{action}}(a) + \nu\right) - \frac{\lambda_p \lambda_a}{\alpha(\lambda_p + \lambda_a + \alpha)} C^{\text{state}}(1) - \frac{\lambda_a}{\alpha}\left(C^{\text{action}}(a) + \nu\right) \\
&\qquad - C^{\text{action}}(a) \\
&\Leftrightarrow \quad \alpha(\lambda_p + \lambda_a + \alpha)\,\nu \leq -\alpha(\lambda_p + \lambda_a + \alpha) C^{\text{action}}(a) \\
&\Leftrightarrow \quad \nu \leq -C^{\text{action}}(a) \quad \text{and} \\
\nu &\overset{(5.14)}{\leq} -\frac{\lambda_p \lambda_a}{\alpha(\lambda_p + \lambda_a + \alpha)} C^{\text{state}}(1) - \frac{\lambda_a}{\alpha}\left(C^{\text{action}}(a) + \nu\right) + \frac{\lambda_a}{\alpha} C^{\text{state}}(1) - C^{\text{action}}(a) \\
&\Leftrightarrow \quad (\lambda_a + \alpha)\,\nu \leq \frac{\lambda_a(\lambda_p + \lambda_a + \alpha) - \lambda_p \lambda_a}{\lambda_a + \lambda_p + \alpha} C^{\text{state}}(1) - (\lambda_a + \alpha) C^{\text{action}}(a) \\
&\Leftrightarrow \quad \nu \leq \frac{\lambda_a}{\lambda_p + \lambda_a + \alpha} C^{\text{state}}(1) - C^{\text{action}}(a),
\end{aligned}$$

which holds if and only if $\nu \leq \min\{-C^{\text{action}}(a), \lambda_a/(\lambda_p + \lambda_a + \alpha) C^{\text{state}}(1) - C^{\text{action}}(a)\}$.

Since we assume $C^{\text{state}}(1), C^{\text{action}}(a) \geq 0$, we have the following conclusions:

1. μ^{pp} is optimal $\Leftrightarrow \nu \geq \lambda_a/(\lambda_p + \alpha) C^{\text{state}}(1) - C^{\text{action}}(a)$.
2. a) If $C^{\text{state}}(1) = 0$, then μ^{ap} is optimal $\Leftrightarrow \nu = -C^{\text{action}}(a)$.
   b) If $C^{\text{state}}(1) > 0$, then μ^{ap} is not optimal for any $\nu \in \mathbb{R}$.
3. μ^{pa} is optimal $\Leftrightarrow -C^{\text{action}}(a) \leq \nu \leq \lambda_a/(\lambda_p + \alpha) C^{\text{state}}(1) - C^{\text{action}}(a)$.
4. μ^{aa} is optimal $\Leftrightarrow \nu \leq -C^{\text{action}}(a)$.

From 1–4, we conclude that the Whittle index has the asserted form and that the project is indexable. □

Note that the summand $-C^{\text{action}}(a)$ appears in $\nu^W(0)$ and in $\nu^W(1)$. It is remarkable that the same term occurs in the original definition of ι in Lemma 5.4.3 if a is optimal.

Proposition 5.4.12. *For a project (P) satisfying Assumption 5.4.10, we have*

$$\iota(0) = 0, \quad \iota(1) = (\lambda_a + \alpha)\left(\frac{C^{\text{state}}(1)}{\alpha} - v^*(1)\right) \quad \textit{and}$$
$$\tilde{\iota}(0) = 0, \quad \tilde{\iota}(1) = \frac{C^{\text{state}}(1)}{\alpha} - v^*(1).$$

where v^* is the optimal value function of (P), and μ^* is an optimal decision rule for (P).

Proof. Since $C^{\text{action}}(a) \geq 0$, it is always optimal to choose action p in state 0 by the proof of Proposition 5.4.11 (since μ^{pp} or μ^{pa} is optimal for $\nu = 0$). Therefore $\iota(0) = \tilde{\iota}(0) = 0$. We have $T_p v^*(1) = C^{\text{state}}(1)/\alpha$. Hence $\tilde{\iota}(0) = C^{\text{state}}(1)/\alpha - v^*(1)$. If a is optimal in state 1, then we have $\iota(s) = (\lambda_a + \alpha)(C^{\text{state}}(1)/\alpha - v^*(1))$. Moreover, we have that p is optimal in state 1 if and only if $T_p v^*(1) = C^{\text{state}}(1)/\alpha = v^*(1)$, so that $C^{\text{state}}(1)/\alpha - v^*(1) = 0$. Together, we have $\iota(1) = (\lambda_a + \alpha)(C^{\text{state}}(1)/\alpha - v^*(1))$. □

Now, assume that the decision maker has to take care of 100 machines each of them satisfying Assumption 5.4.10. The parameters of machine i are denoted by a lower index i, $i = 1, \ldots, 100$. Further, assume that the decision maker can repair at most r machines at the same time. Since $C_i^{\text{action}}(a) \geq 0$, it is optimal to choose the passive action in state 0 for every project. Hence, the decision maker has only to decide which of the machines currently being in state 1 she wants

proj.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
rank	78	74	70	9	68	69	86	99	8	76	82	22	25	75	95	30	60	55	5	62
proj.	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
rank	23	64	89	80	96	10	50	1	84	71	56	31	45	65	97	100	94	43	7	92
proj.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
rank	2	53	42	41	3	4	35	29	54	49	67	91	77	15	73	38	52	79	16	11
proj.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
rank	6	58	51	36	47	21	39	46	59	57	98	88	34	63	90	17	14	81	33	48
proj.	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
rank	12	61	87	26	28	37	19	93	18	44	66	13	40	27	32	83	72	20	85	24

Table 5.1.: Ranks of the projects with respect to ν^W in state 1.

proj.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
rank	75	74	68	50	71	69	75	75	20	75	75	45	21	75	75	34	67	46	1	59
proj.	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
rank	13	66	75	75	75	56	44	2	75	70	64	26	24	60	75	75	75	27	11	75
proj.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
rank	3	38	37	23	30	33	31	6	39	29	61	75	75	9	73	43	51	75	7	40
proj.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
rank	15	54	48	17	36	10	35	16	63	52	75	75	49	57	75	4	41	75	12	42
proj.	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
rank	58	55	75	22	65	14	28	75	8	47	62	25	53	32	19	75	72	5	75	18

Table 5.2.: Ranks of the projects with respect to ι in state 1.

to be repaired. Note that it need not be optimal to repair a machine occupying state 1 since $C_i^{\text{action}}(a)$ might be high in comparison to its benefit when reaching state 0, and therefore, machine i should remain in state 1 forever. The data for the experiments were drawn uniformly from the following sets: $\lambda_{i,p} \in \{1,\dots,50\}$, $\lambda_{i,a} \in \{1,\dots,50\}$, $C_i^{\text{state}}(1) \in \{1,\dots,25\}$ and $C_i^{\text{action}}(a) \in \{1,\dots,12\}$, $i = 1,\dots,100$. Also, some initial state $x_0 \in \{0,1\}^{100}$ was randomly chosen. The discount rate is $\alpha = 0.001$.

For state 1, the ranks of the projects given by the indices ν^W, ι and $\tilde{\iota}$ are illustrated in Tables 5.1–5.3. The rank tells the decision maker with which priority the projects should be worked on if they are in state 1. For the Whittle index $\nu^W(1)$, we have the ranks shown in Table 5.1. So, machine 1 has priority 78, and if machine 28 is in state 1, the index rule given by the Whittle index ν^W chooses to repair machine 19 by all means since $\nu_i^W(0) < \nu_{19}^W(1)$ for all $i = 1,\dots,100$ due to the non-negativity of the model parameters. Here, note that the index rule given by the Whittle index is not only given by Table 5.1, also projects currently being in state 0 have to be considered. The priorities with respect to the heuristic index $\iota(1)$ are given in Table 5.2 and with respect to the modified heuristic index $\tilde{\iota}(1)$ are given in Table 5.3.

As one can see from the tables, there are quite big differences between the project orderings. For instance, project 4 is ranked 9th with respect to the Whittle index, whereas it is ranked 50th with respect to ι and 6th with respect to $\tilde{\iota}$. However, some projects are ranked similarly such as, e. g., project 28 which is ranked 1st under the Whittle index and 2nd under ι and $\tilde{\iota}$. Nevertheless, there remains one major drawback of the Whittle index: one cannot directly deduce

proj.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
rank	75	74	70	6	69	68	75	75	5	75	75	12	26	75	75	32	60	50	1	54
proj.	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
rank	16	62	75	75	75	29	47	2	75	71	58	46	38	64	75	75	75	41	15	75
proj.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
rank	4	53	34	30	3	42	27	14	51	44	61	75	75	10	73	24	45	75	11	18
proj.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
rank	17	49	57	31	37	8	33	35	56	66	75	75	48	59	75	9	20	75	23	43
proj.	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
rank	55	67	75	36	63	22	13	75	19	52	65	21	40	25	28	75	72	7	75	39

Table 5.3.: Ranks of the projects with respect to $\tilde{\iota}$ in state 1.

from it whether the active action is optimal at all for a given project. This comes from the nature of the Whittle index, which gives an optimal policy under the constraint that in discounted average r projects are being worked on at the same time. Hence, even for projects in which it is not optimal to use the active action numbers have to be assigned. The indices ι and $\tilde{\iota}$, of course, include this information, since the single projects are solved separately and $\iota_i(1) = 0$, or $\tilde{\iota}_i(1) = 0$ respectively, if and only if p is optimal in state 1 by Theorem 5.4.6. Note that the highest rank for ι and $\tilde{\iota}$ is 75, meaning that there are 26 machines for which it is never optimal to be repaired.

We used the indices to simulate six different policies:

1. $\mu^{W,r}$ is the original Whittle heuristics: it chooses action a for those r projects having largest indices $\nu_i^W(s_i)$. Here, also projects which currently are in state 0 might be assigned with the active action a.

2. $\tilde{\mu}^{W,r}$ is similar to the Whittle heuristics. It chooses those r projects having largest indices, but it only assigns a to those projects currently being in state 1 and p to those projects being in state 0. Due to the proof of Proposition 5.4.11, action p is optimal in 0 for every machine. Therefore, the heuristics $\tilde{\mu}^{W,r}$ should perform better than $\mu^{W,r}$.

3. $\hat{\mu}^{W,r}$ is also based on the Whittle index. But we use more information about the single projects. Projects in which it is never optimal to use the active action are dropped from consideration. Furthermore, projects which currently are in state 0 are not considered. From the remaining projects, those r projects having largest indices $\nu_i^W(1)$ are assigned with the active action.

4. μ^r is the heuristics derived from the heuristic index ι. The active action a is chosen for those r projects with largest indices $\iota_i(s_i)$ as long as $\iota_i(s_i) > 0$. The passive action p is chosen for the remaining projects. So, only projects currently being in state 1 are considered.

5. $\tilde{\mu}^r$ is derived from the modified heuristic index $\tilde{\iota}$ in the same manner as μ^r. Similar to μ^r, only projects currently being in state 1 are considered.

6. π^r_{random} is a policy which chooses uniformly r projects which should be repaired. This policy is simulated to assess the performances of the above heuristics.

We simulated the 100 projects 1000 times under each policy over a time horizon of length $T_{\text{end}} = 10$ for several $r \in \{0,\dots,90\}$, where every simulation run started in x_0. The results can be seen in Figure 5.4 where the average of the discounted cost up to T_{end} of the respective policies are plotted. For all policies, the standard deviation is less than 100.

The costs for the random policy are the highest of all policies. They decrease first and increase afterwards with a minimum at about $r = 30$. For increasing r, the values of the Whittle heuristics $\mu^{W,r}$, which are marked by the upper dashed line, decrease first and increase again with a minimum at about $r = 25$. Since one has to work on exactly r machines under the Whittle heuristics $\mu^{W,r}$, it might happen that machines have to be worked on for which it is not optimal to work on at some point in time, e. g., machines which currently are in state 0 are worked on although it would be better not to repair those machines. So, additional costs arise. The same holds for $\tilde{\mu}^{W,r}$, which is depicted by the middle dashed line. An explanation for this behaviour is that the Whittle index does not detect projects for which it is not optimal to repair in state 1. The remaining policies perform better for increasing r. For $r > 25$, there is not much change in the simulated costs. The reason is that every machine can be treated optimally at almost any point in time for $r > 25$. Since all machines are independent, a policy which works optimally on every machine at any point in time is optimal by Theorem 4.1.11. From Figure 5.4, one can see that the index policies derived from ι and $\tilde{\iota}$ both perform almost as good as the Whittle heuristics for $r \leq 25$. Nevertheless, policy $\tilde{\mu}^r$ performs better than μ^r. But both policies are good heuristics to derive a suboptimal policy. The policy $\hat{\mu}^{W,r}$ is the best of the considered heuristics. For $r \leq 25$, it performs as good as the policies $\mu^{W,r}$ and $\tilde{\mu}^{W,r}$. For $r > 25$, it performs as good as the heuristics μ^r and $\tilde{\mu}^r$.

In conclusion, if there is no better heuristics available, e. g., because the Whittle index is not computable, then one should use μ^r or $\tilde{\mu}^r$ to obtain a suboptimal policy since both obviously perform better than to randomly choose actions, and both heuristic policies perform almost as good as the Whittle heuristics.

5.4.6. Index-Based Heuristics for the Surveillance Task

In this section, we present how the index ι of section 5.4.4 might be used as the basis of a heuristics for the surveillance task. Assume that the infrastructure consists of n sectors and that the security staff size is r. We propose Algorithm 5.1 in order to derive a heuristics for the surveillance task.

In Algorithm 5.1, $T_{\Sigma_{\text{sub}},0}$ is the one-step operator of the passive action $(0,\dots,0)$ with respect to the subinfrastructure Σ_{sub}. The algorithm splits the original infrastructure into subinfrastructures of size m. These subinfrastructures are

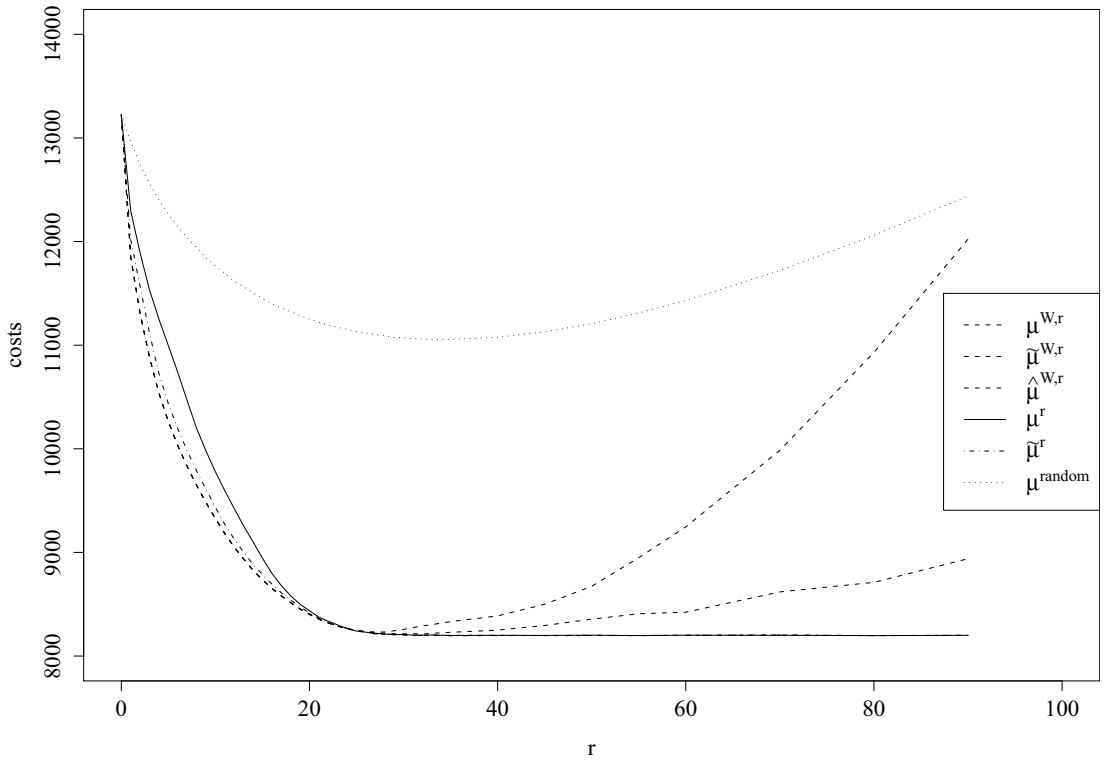


Figure 5.4.: Performance of the six simulated policies.

Algorithm 5.1 Computation of the heuristics for the surveillance task.

Require: parametrization of the infrastructure Σ according to chapter 3, $m \geq r$

1: **for** all subinfrastructures $\Sigma_{\text{sub}} = \{\sigma_1, \ldots, \sigma_m\} \subset \Sigma$ of size m **do** {off-line computations}

2: determine from the parametrization of Σ restricted to Σ_{sub} the optimal value function $v^*_{\Sigma_{\text{sub}}}$ and an optimal decision rule $\mu^*_{\Sigma_{\text{sub}}}$ of Σ_{sub}

3: **for** all $s_{\text{sub}} \in G^m$ **do**

4: determine $\iota_{\Sigma_{\text{sub}}}(s_{\text{sub}}) = \left(\lambda_{\Sigma_{\text{sub}}}(s_{\text{sub}}, \mu^*_{\Sigma_{\text{sub}}}(s_{\text{sub}})) + \alpha\right)\left(T_{\Sigma_{\text{sub}},0} v^*_{\Sigma_{\text{sub}}}(s_{\text{sub}}) - v^*_{\Sigma_{\text{sub}}}(s_{\text{sub}})\right)$

5: **end for**

6: **end for**

7: **for** $s \in S$ **do** {on-line computations}

8: determine $\Sigma^*_{\text{sub}}(s)$ such that

$$\iota_{\Sigma^*_{\text{sub}}(s)}\left(s_{\Sigma^*_{\text{sub}}(s)}\right) = \max_{\substack{\Sigma_{\text{sub}} \subset \Sigma: \\ \#\Sigma_{\text{sub}} = m}} \left\{\iota_{\Sigma_{\text{sub}}}(s_{\Sigma_{\text{sub}}})\right\},$$

where $s_{\Sigma_{\text{sub}}}$ is the restriction of s to the sectors of $\Sigma_{\text{sub}} \subset \Sigma$

9: define

$$\mu_{\text{heur}}(s)(\sigma) := \begin{cases} \mu^*_{\Sigma^*_{\text{sub}}(s)}(s_{\Sigma^*_{\text{sub}}(s)})(\sigma), & \text{if } \sigma \in \Sigma^*_{\text{sub}}(s) \\ 0, & \text{else} \end{cases}, \quad \sigma \in \Sigma$$

10: **end for**

11: **return** μ_{heur}

treated as if they were independent projects. In short, for a given state $s \in S$ of the original infrastructure, the heuristics chooses a subinfrastructure $\Sigma^*_{\text{sub}}(s)$ with maximal index and assigns the respective active elementary actions to the original infrastructure. For sectors which do not lie in $\Sigma^*_{\text{sub}}(s)$, the passive elementary action 0 is chosen.

Remark 5.4.13.

1. Note that Algorithm 5.1 requires $m \geq r$ so that the entire security staff can be assigned to the subinfrastructures. Therefore, subinfrastructures of size m have to be computable in adequate time. In our examples, m should not exceed five.

2. The number of considered subinfrastructures might also be a limiting factor when using this heuristics. This number is given by $\binom{n}{m}$.

3. In Algorithm 5.1, the computation of the indices for the subinfrastructures in lines 1–6 should be done in advance before implementing the decision support system. For the current state $s \in S$, the heuristic suboptimal action $\mu_{\text{heur}}(s)$ can be determined on-line during operation of the decision support system so that the entire heuristics μ_{heur} need not be known in advance. So, the heuristic decision rule μ_{heur} need not be stored on a hard disk drive, but only the indices and optimal decision rules for the subinfrastructures.

4. Because of its structure, the heuristics makes several errors. First of all, dependencies from a subinfrastructure to its complement sectors are not considered in any way. Furthermore, the subinfrastructures are considered as independent projects, which is not the case due to the dependencies. Moreover, the index ι itself is also only a heuristics to find the most beneficial subinfrastructure to which the resources should be allocated if the considered subinfrastructures were independent.

5. If r is larger than any m for which the surveillance task of subinfrastructures of size m is computable, one could try to find other heuristics. We propose two simple heuristics which are both based on ι.

   a) At first, we solve the surveillance tasks for all one-sector subinfrastructures. For some given $s \in S$, one could then define a heuristic action by assigning the optimal action of those r subinfrastructures $\{\sigma\} \subset \Sigma$ which have the highest indices $\iota_{\{\sigma\}}(s(\sigma))$. This approximation method does not take into account any dependencies between the sectors and treats the sectors to be independent of each other. For $r = n$, i. e., there are no resource restrictions, the resulting error of this heuristics is considered in section 4.2.

   b) At first, we solve the surveillance task for all m-sector subinfrastructures with m resources, where m is chosen such that the respective surveillance tasks are exactly computable. Now, let $s \in S$ be given. First, we take the subinfrastructure with maximal index and assign the respective elementary actions to the heuristic decision rule. Next, we take the subinfrastructure with second-largest index. There are several cases: if this subinfrastructure chooses an active elementary action in a sector we have already assigned another active elementary action, then we drop this sector. We assign the optimal active elementary action to the respective sector if enough resources are still available. If one can only assign some active elementary actions to the heuristic decision rule due to the resource restriction, ties are broken by the number of dependencies or by a random choice. In this manner, we go through every subinfrastructure in descending order of the indices. We stop if we have assigned an active elementary action to r subinfrastructures. It might be the case that there are some resources left. In this case, we consider all one-sector subinfrastructures to which no active elementary action has already been assigned. We solve the associated surveillance tasks. Again, we consider the subinfrastructures in descending order of the respective indices and assign the respective active elementary action to the heuristics until r active elementary actions are assigned to the heuristic decision rule or until we have gone through all one-sector subinfrastructures, in which case not all resources are needed for the heuristic decision rule. In this way, dependencies of the infrastructure are considered within the subinfrastructures.

   Other heuristics may be derived by considering several subinfrastructures with different sizes and different resource capacities and by composing the heuristics from optimal actions of the subinfrastructures with highest indices.

5.4.7. Numerical Experiments

In this section, we investigate the behaviour of the heuristics presented in the preceding section. We do not only consider the heuristics μ_{heur} but also an ad hoc heuristics $\mu_\bigstar$, which is marked with $\bigstar$ in the subsequent tables. In the following surveillance tasks, we have included elementary action 2, i. e., "Inspection walk," of the accompanying example of chapter 3. Since an accomplished elementary action 2 leaves the respective sector in state 0 and also decreases the threat level of dependent sectors, we define $\mu_\bigstar$ in the following way: in the subsequent examples, there are threat events $e^1_\sigma \in \mathscr{E}(\sigma)$, $\sigma \in \Sigma$, which model the destruction of σ. For $r = 1$, $\mu_\bigstar$ assigns elementary action 2 to the sector which has the highest threat level. If two or more sectors have the highest threat level, 2 is assigned to the sector $\sigma \in \Sigma$ with

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	2.2 %	9.7 %	32.5 %	14.0 %
1	2	0.6 %	3.3 %	17.7 %	9.4 %
1	3	0.1 %	1.0 %	10.1 %	5.3 %
1	★	1.0 %	3.8 %	36.4 %	16.3 %
2	2	1.2 %	5.1 %	14.8 %	9.0 %
2	3	0.7 %	2.6 %	12.0 %	7.7 %
2	★	1.2 %	6.5 %	30.3 %	17.0 %

Table 5.4.: Results for the accompanying example of chapter 3.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	2.2 %	11.1 %	43.4 %	15.2 %
1	2	1.0 %	4.6 %	20.2 %	9.8 %
1	3	0.4 %	1.3 %	7.8 %	4.4 %
1	★	1.0 %	3.8 %	36.4 %	16.3 %
2	2	2.4 %	7.5 %	20.6 %	19.0 %
2	3	1.2 %	3.8 %	19.4 %	8.9 %
2	★	1.2 %	6.5 %	30.3 %	17.0 %

Table 5.5.: Results for the accompanying example of chapter 3 with respect to index $\tilde{\iota}$.

maximal destruction cost $C_{e^1_\sigma}$. For $r = 2$, it assigns elementary action 2 to the sectors which have the two highest threat levels. Again, ties are broken by considering the respective destruction costs $C_{e^1_\sigma}$ and assigning elementary action 2 to the sectors with highest destruction costs.

At first, we consider the accompanying example of chapter 3 with security staff sizes $r = 1,2$. The results can be obtained from Table 5.4.

In Table 5.4 and in the following tables, we give the minimal and the maximal relative error between the heuristic value functions $v^{\mu_{\text{heur}}}$ and the optimal value function v^* of the respective surveillance tasks, i. e., $\min_{s\in S}\{v^{\mu_{\text{heur}}}(s)/v^*(s) - 1\} \cdot 100\,\%$ and $\max_{s\in S}\{v^{\mu_{\text{heur}}}(s)/v^*(s) - 1\} \cdot 100\,\%$ respectively. Furthermore, the average relative error computed by $[\sum_{s\in S}[(v^{\mu_{\text{heur}}}(s) - v^*(s))/v^*(s)]/|S|] \cdot 100\,\%$ is given. The relative error is given by $\|v^{\mu_{\text{heur}}} - v^*\|_\infty / \|v^*\|_\infty \cdot 100\,\%$.

For $r = 2$, Figure 5.5 illustrates the heuristic decision rule for a two-sectors approximation, i. e., $m = 2$, and Figure 5.6 shows the heuristic decision rule for a three-sectors approximation, i. e., $m = 3$. A comparison shows that the heuristic decision rules have roughly the same structure as the optimal decision rule depicted above in Figure 3.2. The heuristics μ_{heur} works well for this example for all m. Moreover, the simple heuristics $\mu_\star$ is quite good, too. Note that there are decision rules which are very bad for this very surveillance task as can be seen in section 5.3.3.

In Table 5.5, the same infrastructure is considered. But this time, $\tilde{\iota}$ as defined in Remark 5.4.7 is used as an index in lines 4 and 8 of Algorithm 5.1. This heuristics performs worse than the original heuristics. Therefore, the heuristics μ_{heur} is based on ι.

Now, we consider an infrastructure consisting of five sectors since we are able to solve the respective surveillance tasks exactly. The parameter set is stated in section B.2. We computed the value functions for the heuristics given in the preceding section for $r = 1,2$ and $r \leq m \leq 4$. Later, we vary these parameters in order to see how the quality of the heuristics is influenced by the particular models of the surveillance task.

One can see that the heuristics μ_{heur} works quite well for $r = 1$. It is remarkable that the one-sector heuristics is better than the two-sectors and the three-sectors heuristics. Also the simple heuristics $\mu_\star$ is quite good for $r = 1$. But for $r = 2$, the performance of the heuristics improves as m is getting larger. In this case, the heuristics μ_{heur} and $\mu_\star$ both perform very well.

The next experiment includes basically the same infrastructure, except that we have changed the dependencies between

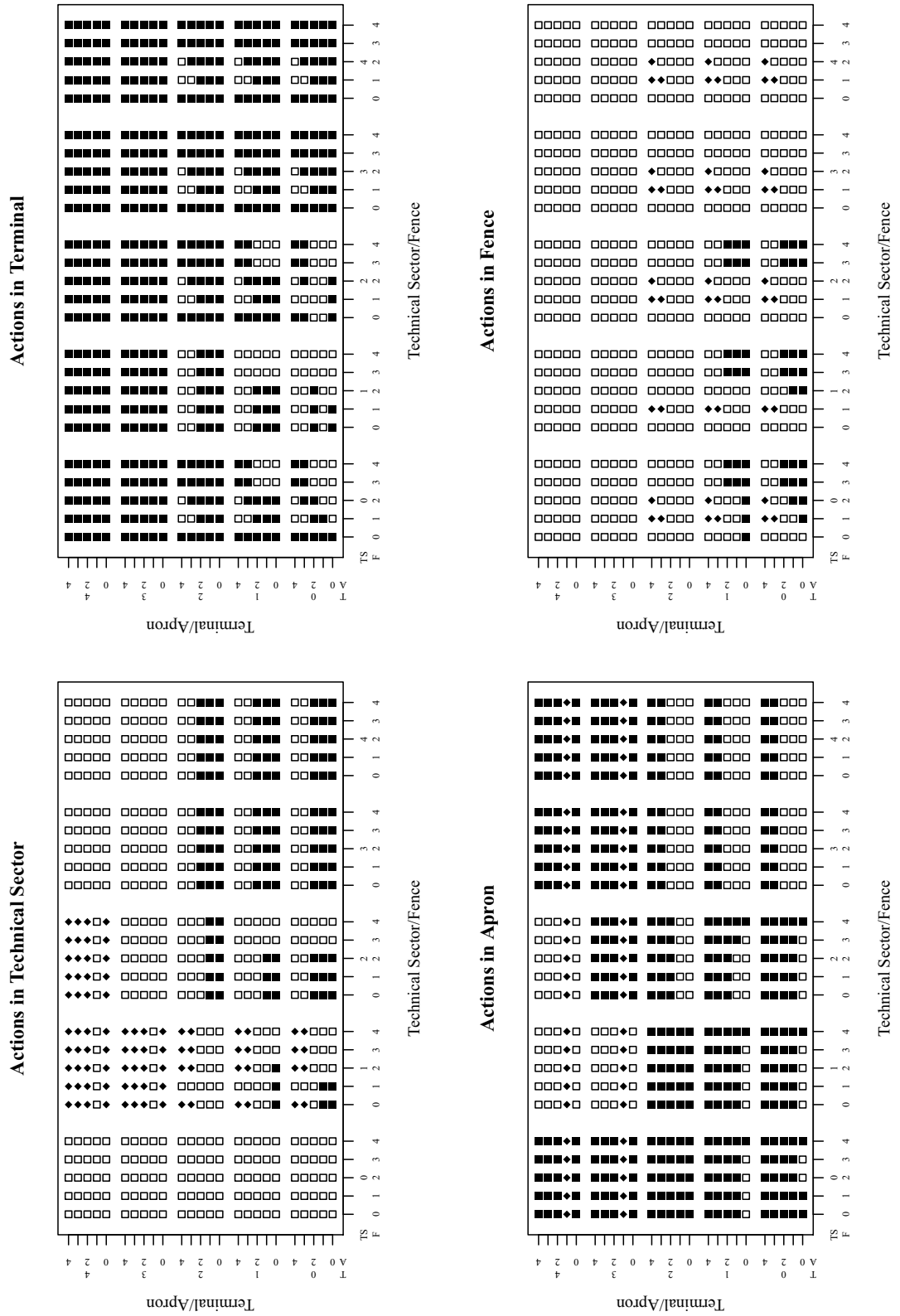


Figure 5.5.: Heuristic decision rule for the accompanying example for $r = 2$ and $m = 2$.

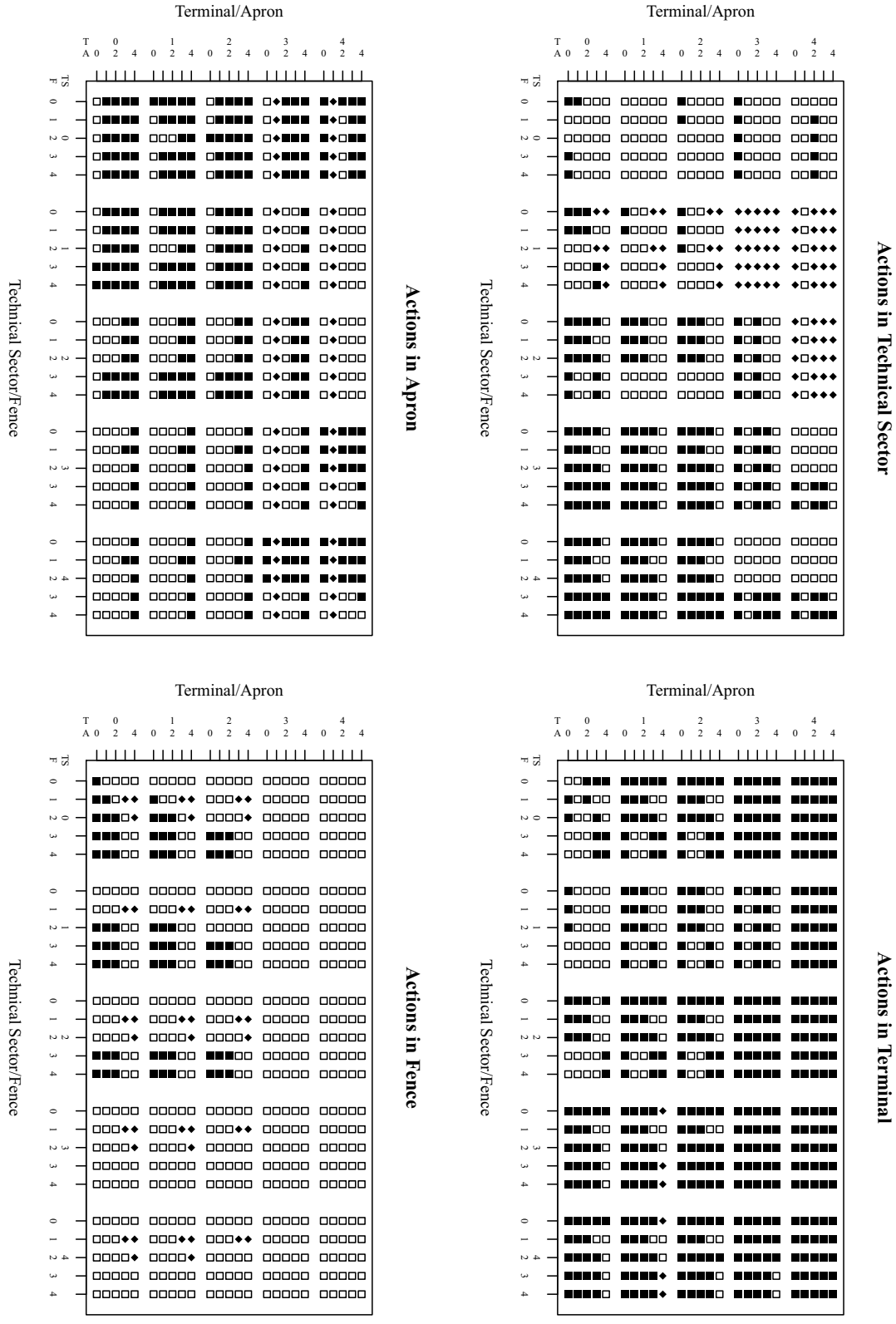


Figure 5.6.: Heuristic decision rule for the accompanying example for $r = 2$ and $m = 3$.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	6.8 %	30.3 %	93.9 %	45.8 %
1	2	20.0 %	57.9 %	114.0 %	60.2 %
1	3	13.9 %	39.7 %	81.9 %	41.7 %
1	4	4.7 %	14.7 %	45.2 %	21.7 %
1	★	6.7 %	28.8 %	93.4 %	45.5 %
2	2	2.1 %	17.2 %	46.0 %	23.7 %
2	3	1.0 %	11.5 %	41.5 %	24.9 %
2	4	0.3 %	4.8 %	26.5 %	11.5 %
2	★	0.7 %	14.5 %	83.1 %	26.5 %

Table 5.6.: Results for the original numerical experiment.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	1.6 %	83.5 %	423.0 %	159.2 %
1	2	0.6 %	35.5 %	120.4 %	74.7 %
1	3	0.5 %	31.9 %	80.7 %	50.6 %
1	4	0.2 %	12.2 %	60.3 %	39.6 %
1	★	1.1 %	35.9 %	112.7 %	68.8 %
2	2	2.2 %	22.8 %	77.6 %	35.0 %
2	3	1.8 %	13.1 %	41.7 %	21.9 %
2	4	0.5 %	6.0 %	30.6 %	18.6 %
2	★	1.3 %	20.8 %	89.2 %	22.3 %

Table 5.7.: Results for the numerical experiment with very few dependencies.

the sectors. Here, we assume that the infrastructure has very few dependencies such that

$$N = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

The results for this infrastructure are listed in Table 5.7. Again, the performance is satisfactory for the heuristics for $r = 1$. But the performance of the heuristics is even better if the security staff size is two. In both cases, the performance improves with increasing m.

In the next experiment, we have much more dependencies between the sectors, but all other parameters are the same as above: here, we have

$$N = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

The results for this infrastructure are given in Table 5.8. For $r = 1$, the heuristics perform rather bad. However, μ_{heur} is clearly better than $\mu_\star$ for $m \geq 2$. Whereas for $r = 2$, the heuristics is very good for $m \geq 3$, and definitely beats the simple heuristics $\mu_\star$ even for $m = 3$. Again by increasing m, the performance of the heuristics μ_{heur} improves.

Now, we consider the original five-sectors infrastructure. But, we assume that the costs of destruction are given by

$$C_{e^1_{\sigma_1}} = 100{,}000, \quad C_{e^1_{\sigma_2}} = 10{,}000{,}000, \quad C_{e^1_{\sigma_3}} = 500{,}000, \quad C_{e^1_{\sigma_4}} = 500{,}000, \quad C_{e^1_{\sigma_5}} = 500{,}000.$$

So, one sector is extremely more expensive than the others. The results are presented in Table 5.9. For $r = 1$ and $r = 2$, the simple heuristics $\mu_\star$ works very well. But at least for $r = 2$ and $m \geq 2$, the performance of the heuristics μ_{heur} is very good, too.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	109.3 %	414.7 %	743.8 %	418.4 %
1	2	50.1 %	175.0 %	311.1 %	208.4 %
1	3	44.3 %	166.6 %	273.3 %	211.0 %
1	4	39.6 %	147.7 %	224.8 %	196.4 %
1	★	108.0 %	412.3 %	670.6 %	418.3 %
2	2	12.4 %	94.8 %	189.5 %	102.4 %
2	3	2.2 %	17.6 %	52.5 %	26.2 %
2	4	0.8 %	9.5 %	27.0 %	16.4 %
2	★	8.3 %	82.8 %	209.2 %	89.7 %

Table 5.8.: Results for the numerical experiment with many dependencies.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	11.9 %	44.0 %	140.6 %	52.7 %
1	2	18.3 %	54.7 %	168.3 %	66.3 %
1	3	20.7 %	53.7 %	127.0 %	65.3 %
1	4	10.3 %	31.2 %	67.4 %	43.9 %
1	★	4.8 %	24.0 %	84.1 %	39.8 %
2	2	5.3 %	42.8 %	131.2 %	48.1 %
2	3	1.4 %	16.7 %	69.3 %	26.5 %
2	4	0.5 %	6.9 %	31.0 %	17.4 %
2	★	1.0 %	15.8 %	98.4 %	43.5 %

Table 5.9.: Results for the numerical example with one very expensive sector.

In the next experiment, the rates of the actions depend on the sectors. The remaining parameters are the same as in the first experiment. We have

$$\lambda_1(\sigma_1) = 30, \quad \lambda_1(\sigma_3) = 10, \quad \lambda_1(\sigma_3) = 15, \quad \lambda_1(\sigma_4) = 20, \quad \lambda_1(\sigma_5) = 40,$$
$$\lambda_2(\sigma_1) = 10, \quad \lambda_2(\sigma_3) = 5, \quad \lambda_2(\sigma_3) = 5, \quad \lambda_2(\sigma_4) = 20, \quad \lambda_2(\sigma_5) = 15.$$

The results are given in Table 5.10. It is obvious that the quality improves when m increases. For $r = 1$, the ad hoc heuristics is bad as well as μ_{heur} for $m = 1$. Furthermore, the heuristics μ_{heur} performs better than the simple heuristics $\mu_\star$ for $r = 2$ and $m \geq 3$.

In the next experiment, we have set the discount rate to $\alpha = 10^{-6}$, but all remaining parameters are the same as in the first five-sectors experiment. The results can be seen in Table 5.11. Surprisingly, the performance does not improve for $r = 1$ when m is increased. The simple heuristics $\mu_\star$ is quite good, but the heuristics with respect to $m = 4$ is even better than $\mu_\star$. For $r = 2$, all heuristics seem to be very fine. Nevertheless, the four-sectors heuristics is the best of the considered heuristics.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	69.0 %	249.0 %	723.7 %	174.9 %
1	2	25.7 %	90.5 %	333.2 %	71.3 %
1	3	7.1 %	30.1 %	179.4 %	37.7 %
1	4	0.6 %	6.6 %	82.1 %	24.9 %
1	★	72.6 %	239.4 %	714.3 %	172.0
2	2	9.4 %	63.1 %	200.2 %	55.3 %
2	3	2.6 %	22.0 %	88.6 %	36.5 %
2	4	0.9 %	13.1 %	69.8 %	25.0 %
2	★	5.0 %	41.8 %	203.2 %	61.6 %

Table 5.10.: Results for the numerical example with varied action rates.

r	m	Minimal relative error	Average relative error	Maximal relative error	Relative error
1	1	6.9 %	7.0 %	7.0 %	7.0 %
1	2	20.3 %	20.3 %	20.4 %	20.4 %
1	3	14.2 %	14.2 %	14.2 %	14.2 %
1	4	4.7 %	4.8 %	4.8 %	4.8 %
1	★	6.8 %	6.8 %	6.9 %	6.8 %
2	2	2.1 %	2.1 %	2.1 %	2.1 %
2	3	1.0 %	1.0 %	1.1 %	1.1 %
2	4	0.3 %	0.3 %	0.3 %	0.3 %
2	★	0.7 %	0.7 %	0.7 %	0.7 %

Table 5.11.: Results for the numerical example with discount rate $\alpha = 10^{-6}$.

In conclusion, the heuristics μ_{heur} is more robust in comparison with the simple heuristics $\mu_{\bigstar}$. Especially, if the infrastructure has many dependencies or if the rates of the elementary actions differ for the sectors, the performance of μ_{heur} is better than $\mu_{\bigstar}$. It is beneficial to choose m as large as possible in general. However, Figures 5.5 and 5.6 suggest that the heuristics μ_{heur} tends to assign elementary actions to sectors with high destruction costs. So, the actions proposed by the heuristics are plausible to the decision maker. Therefore, the heuristics μ_{heur} might be used as the basis of a decision support system.

5.4.8. Memory Requirements

Using the heuristics of section 5.4.6, the memory requirements are significantly reduced. To see this, we consider an infrastructure of size n where the security staff size is r. If the heuristics works with an m-sectors approximation with $r \leq m$, then $\binom{n}{m}$ subinfrastructures have to be considered. To determine the heuristic decision rule, it is sufficient to store the indices and the optimal decision rules for all subinfrastructures of size m. This approach requires less memory since the state spaces are lower in dimension than the underlying infrastructure.

Example 5.4.14. For a twelve-sectors infrastructure with $g_{\max} = 4$ which is approximated by subinfrastructures of size five, storing an optimal decision rule for a subinfrastructure requires $5^5 \cdot 5 \cdot 8\,\text{bits} \approx 15.6\,\text{kB}$ of memory for each subinfrastructure. Assuming that the indices for one subinfrastructure are stored in a text-file at a length of about 16 digits, then it takes about $5^5 \cdot 16 \cdot 8\,\text{bits} = 50.0\,\text{kB}$ to store the indices file for one subinfrastructure. All in all, the memory requirements are $\binom{12}{5}(15.6 + 50.0)\,\text{kB} \approx 52.0\,\text{MB}$ meaning that the memory requirements are reduced by about 95 %.

6. Risk Measures for the Surveillance Task

In chapter 3, we modelled the dynamics of threat of a critical infrastructure in terms of a CMDP. Then we concentrated on the expected total discounted cost under deterministic stationary policies since it is known that there exists an optimal policy of this kind according to Theorem 2.3.4. In the context of the surveillance task, expected total discounted cost measures the expected discounted frequency of the occurrence of threat events weighted with their associated costs. Discounting causes a weight on the time of occurrence of threat events such that the earlier a threat event occurs the more influence it has on the total discounted cost. The objective in the previous chapters was to find a decision rule which minimizes the expected total discounted cost. Now, the question arises whether this criterion is appropriate for the surveillance task or whether the objective criterion should be modified in order to comprise the infrastructure owner's or decision maker's preferences.

Certainly, the primary goal of a decision maker who is in charge of a critical infrastructure is to minimize the occurrence of threat events which would heavily harm the infrastructure. For this purpose, the decision maker might accept the occurrence of several threat events which have a minor impact on the infrastructure as long as disastrous threat events are prevented in this way. Therefore, the decision maker is risk-averse. But this risk aversion is not included within the framework of the expected total discounted cost, where all threat events are just weighted with their probability of occurrence and corresponding costs, no matter if they are disastrous or if they have a low impact on the infrastructure. In order to formalize the decision maker's preferences or risk aversion, appropriate risk measures might be defined. At first, we introduce a risk measure which is appropriate for the surveillance task. In the risk aversion framework, we seek policies which minimize the risk measure of the total discounted cost in chapter 7 and for the average cost in chapter 8 for some given initial state.

6.1. Risk Measures

In this section, we shortly summarize the concept of risk measures and its meanings in financial applications. We consider the case of integrable random variables only.

Definition 6.1.1. Let $L^1(\Omega,\mathscr{A},P)$ be the space of all real-valued integrable random variables on some probability space $(\Omega,\mathscr{A},P)$. A *risk measure* ρ is a function $\rho : L^1(\Omega,\mathscr{A},P) \to \mathbb{R}$.

So, a risk measure assigns a real value to each integrable random variable. In particular, the expectation and the variance, if it exists, or any other characteristics of random variables X are risk measures. In finance, according to (McNeil et al., 2005), p. 239, $\rho(X)$ should be interpreted as the amount of capital that should be added to a position with a loss given by X, so that the position becomes acceptable to an external risk controller. If $\rho(X) \leq 0$, then the position X is acceptable without injection of capital and if $\rho(X) < 0$, then capital may even be withdrawn. There are several properties that a reasonable risk measure should satisfy in order to adequately measure risk in financial applications. To be consistent with the following, we assume that costs are measured by positive values and that rewards are measured by negative values respectively.

Definition 6.1.2. A risk measure ρ is a *convex risk measure* if it satisfies the following properties:

1. For all $X_1, X_2 \in L^1(\Omega,\mathscr{A},P)$ such that $X_1 \leq X_2$, we have $\rho(X_1) \leq \rho(X_2)$ *(monotonicity).*
2. For all $X \in L^1(\Omega,\mathscr{A},P)$ and $c \in \mathbb{R}$, we have $\rho(X+c) = \rho(X) + c$ *(translation equivariance).*
3. For all $X_1, X_2 \in L^1(\Omega,\mathscr{A},P)$ and $\lambda \in (0,1)$, we have $\rho(\lambda X_1 + (1-\lambda)X_2) \leq \lambda\,\rho(X_1) + (1-\lambda)\,\rho(X_2)$ *(convexity).*

Definition 6.1.3. A convex risk measure ρ is a *coherent risk measure* if ρ is *positively homogeneous,* i. e., for every $X \in L^1(\Omega,\mathscr{A},P)$ and every $\lambda > 0$, we have $\rho(\lambda X) = \lambda\,\rho(X)$.

Definition 6.1.4. A risk measure ρ is *subadditive* if $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$ for all $X_1, X_2 \in L^1(\Omega,\mathscr{A},P)$.

Convex and coherent risk measures are extensively studied in the literature (cf. (Artzner et al., 1999; McNeil et al., 2005)). Examples for coherent risk measures include the expectation and the average value-at-risk, which is considered in section 6.2.2. In contrast, value-at-risk, which is considered in section 6.2.1, is not even a convex risk measure. The properties in the above definitions have their respective economical interpretations. Monotonicity is clear: if one risk

is higher than another, then more capital has to be added in the first case so that it is acceptable to an external risk controller. Translation equivariance means that when adding $\rho(X)$ to a position X, which corresponds to subtracting it from the position X, then the position $X-\rho(X)$ is acceptable to the risk controller since $\rho(X-\rho(X))=0$. Convexity and subadditivity reflect the idea that diversification should be rewarded by a reduced risk, i. e., splitting risks should be rewarded.

Positive homogeneity is the most controversially discussed property. It means that risk is proportional to the position. In (Föllmer and Schied, 2004), this property is questioned since risk might depend in a non-proportional manner.

A remarkable relation of risk measures is the following: if a risk measure is positively homogeneous, then it satisfies the property of convexity if and only if it is subadditive. For a much more detailed introduction to risk measures used in finance, we refer to (McNeil et al., 2005).

In the surveillance task, things are different from the financial framework. This lies in the fact that in finance, decision makers have to arrange a portfolio consisting of several financial instruments. Each financial instrument comes with a certain risk. Therefore, risk measure theory based on the above meaningful definitions is needed in order to adequately measure risk. In contrast, we only have one infrastructure which has to be taken care of in the surveillance task so that, e. g., convexity is not a meaningful property of measuring risk in this case.

Nevertheless, some of the above properties are meaningful in the surveillance task. We measure threat by the total discounted cost that the dynamics of threat generates. Then monotonicity is a natural property of an adequate risk measure: consider two infrastructures which are modelled according to chapter 3 with the only difference being that the costs of the first infrastructure are lower than the costs of the second infrastructure. Then the total discounted cost of the first infrastructure is lower than the total discounted cost of the second one. We would assume that risk of the second infrastructure is higher than risk of the first one since it is more expensive. Hence, monotonicity should be a property of an adequate risk measure. In a similar manner, translation equivariance is also a reasonable property of an adequate risk measure. The notion of diversification cannot be applied to the surveillance task since the decision maker cannot invest in several infrastructures. She has to take care of the one infrastructure she is in charge of and has to provide safety and security. The concept of positive homogeneity could be applied to the surveillance task. By multiplying the costs of the threat events with some factor, risk might be increased by the same factor.

So, the concept of a reasonable, i. e., convex or coherent, risk measure for financial applications cannot be transferred easily to the surveillance task. Therefore, we consider two risk measures from scratch and examine whether their respective definitions are meaningful in the framework of the surveillance task.

6.2. Two Risk Measures for the Surveillance Task

As we have seen, the expectation does not resemble the decision maker's preferences since her main goal is to prevent disastrous threat events. Therefore, let us consider two more risk measures which are popular in finance and examine whether these risk measures are applicable in the surveillance task. The two risk measures considered here are the value-at-risk and the average value-at-risk.

6.2.1. Value-at-Risk for the Surveillance Task

A desirable property of a "good" policy, i. e., well suited for the surveillance task from the decision maker's point of view, might be that the value of the total discounted cost that would be exceeded with a probability of $(1-\tau)\cdot 100\,\%$ under the same policy is minimized for some given $\tau\in(0,1)$. We call τ the *confidence level.* In practice, one would choose, e. g., $\tau=0.95$ or $\tau=0.99$. This approach would lead to the objective criterion of minimizing the τ-quantile, i. e., the value-at-risk at level τ, of the total discounted cost.

Definition 6.2.1. Let X be a real-valued random variable on some probability space $(\Omega,\mathscr{A},P)$, and let $\tau\in(0,1)$. The *value-at-risk of X at confidence level τ,* denoted by $\mathrm{V@R}_\tau(X)$, is defined as

$$\mathrm{V@R}_\tau(X) := F_X^{-1}(\tau) := \inf\{x\in\mathbb{R}\,|\,P(X\le x)\ge\tau\},$$

where F_X^{-1} is the *quantile function of X*.

In Definition 6.2.1, it has to be taken care that costs are measured by positive numbers and rewards are measured by negative numbers respectively. In case that costs are measured by negative numbers, one defines the value-at-risk of X at level τ by $\mathrm{V@R}_\tau(X) := F_{-X}^{-1}(\tau)$ so that indeed the distribution of the cost is considered.

In general, value-at-risk is not a convex risk measure. Value-at-risk is monotonic, translation equivariant and positively homogeneous. But it is not subadditive in general, which is one point of criticism, e. g., of (Artzner et al., 1999). However,

if the risks are elliptically distributed, then value-at-risk is subadditive as noted in (McNeil et al., 2005). Hence, if the risks are elliptically distributed, then the value-at-risk is a coherent risk measure.

Applying value-at-risk to the surveillance task, a policy which minimizes the value-at-risk would reduce the probability that very expensive, i. e., disastrous, threat events occur to some extent. The major disadvantage of using the value-at-risk approach is that it does not take into account that the total discounted cost might exceed the optimal 0.95-quantile. So, the 0.95-quantile is minimized by some policy, assuming such a policy exists, but its value does not give any information about the total discounted cost which exceed it. This is another point criticism of the value-at-risk as, e. g., (McNeil et al., 2005), p. 38, note. This is illustrated in a small example: consider two random variables X_1, X_2 representing costs with $P(X_1 = 0) = P(X_2 = 0) = 0.8$ and $P(X_1 = 100) = P(X_2 = 1000) = 0.2$. Then we have $\mathrm{V@R}_{0.8}(X_1) = \mathrm{V@R}_{0.8}(X_1) = 0$. But since there is a probability of 0.2 that X_2 is 1000 whereas X_1 is 100 with the same probability, a decision maker would prefer X_1 over X_2. This demonstrates that very catastrophic threat events might still occur with rather high probability. For this reason, the value-at-risk criterion should not be used to model the decision maker's preferences.

Nevertheless, value-at-risk is used as a risk measure in several fields of application. Apart from finance, where value-at-risk took its start, there are, e. g., inventory control (cf. (Luciano et al., 2003; Tapiero, 2003)), foodservice industry (cf. (Sanders and Manfredo, 2002)) and layers of protection analysis (cf. (Fang et al., 2007)). In (Luciano et al., 2003) and in (Sanders and Manfredo, 2002), value-at-risk is taken as a describing statistical characteristic of a probability distribution. Whereas in (Tapiero, 2003) and in (Fang et al., 2007), minimizing the value-at-risk is used as the objective.

6.2.2. Average Value-at-Risk for the Surveillance Task

To avoid the disadvantage of not including the $(1-\tau)\cdot 100\,\%$ highest outcomes of the total discounted cost when using value-at-risk as the objective criterion, one could try to find a policy which minimizes the average of the $(1-\tau)\cdot 100\,\%$ highest possible outcomes of the total discounted cost. This procedure leads to the average value-at-risk criterion at level τ for the total discounted cost.

Definition 6.2.2. Let $X \in L^1(\Omega, \mathscr{A}, P)$ and $\tau \in (0,1)$. The *average value-at-risk of X at confidence level τ,* denoted by $\mathrm{AV@R}_\tau(X)$, is defined as

$$\mathrm{AV@R}_\tau(X) := \frac{1}{1-\tau}\int_\tau^1 \mathrm{V@R}_t(X)\,dt.$$

This definition follows (Bäuerle and Mundt, 2005). Average value-at-risk is sometimes called expected shortfall or conditional value-at-risk. But as mentioned in (Bäuerle and Mundt, 2005), these terms are also used for several other risk measures. Therefore, the reader has to be careful which risk measure is meant in a particular text. From the definition, it can be seen that the average value-at-risk is exactly the mean of the $(1-\tau)\cdot 100\,\%$ highest costs that might occur. Alternatively, the average value-at-risk can be defined as the mean of the τ-tail distribution of X like in (Rockafellar and Uryasev, 2002), where the τ-tail distribution of X is defined by the cumulative distribution function

$$F_{\tau\text{-tail}}(x) = \begin{cases} 0, & \text{if } x < \mathrm{V@R}_\tau(X) \\ \frac{P(X\le x)-\tau}{1-\tau}, & \text{if } x \ge \mathrm{V@R}_\tau(X) \end{cases}, \quad x \in \mathbb{R}.$$

The cumulative distribution function of the τ-tail distribution is zero below $\mathrm{V@R}_\tau(X)$, and above $\mathrm{V@R}_\tau(X)$, it is essentially the same as the cumulative distribution function of X, but it is stretched with the factor $1/(1-\tau)$. In Figure 6.1, $\mathrm{V@R}_\tau(X)$ and $\mathrm{AV@R}_\tau(X)$ are illustrated for some random variable X with cumulative distribution function F_X. As long as $\mathrm{V@R}_\tau(X)$ is non-negative, the average value-at-risk is the appropriately normalized hatched area illustrated in Figure 6.1.

The average value-at-risk is a common risk measure which enjoys great popularity in finance. This is because the average value-at-risk is a coherent risk measure. This result is for instance presented in (Rockafellar and Uryasev, 2002). This is due to the fact that probability atoms are split appropriately. Risk measures that pursue the same idea as the average-value-at-risk, that is to take the mean of the loss above some level in a certain sense, like upper average value-at-risk, defined by $E[X \mid X > \mathrm{V@R}_\tau(X)]$, or lower average value-at-risk, defined by $E[X \mid X \ge \mathrm{V@R}_\tau(X)]$ (cf. (Rockafellar and Uryasev, 2002)), do not split a present probability atom at $\mathrm{V@R}_\tau(X)$, and therefore, they are not coherent risk measures. For our purpose, it is sufficient that the average value-at-risk at level τ also takes into account values which are greater than the value-at-risk at level τ.

An interesting result concerning the average value-at-risk is provided in Theorem 10 of (Rockafellar and Uryasev, 2002). Its proof can also be found there. In (Rockafellar and Uryasev, 2002), the average value-at-risk is studied in very detail and an abundance of its properties is given. Furthermore, it is compared with the lower and the upper average value-at-risk. This theorem is a fundamental building block in establishing the results of the following chapters.

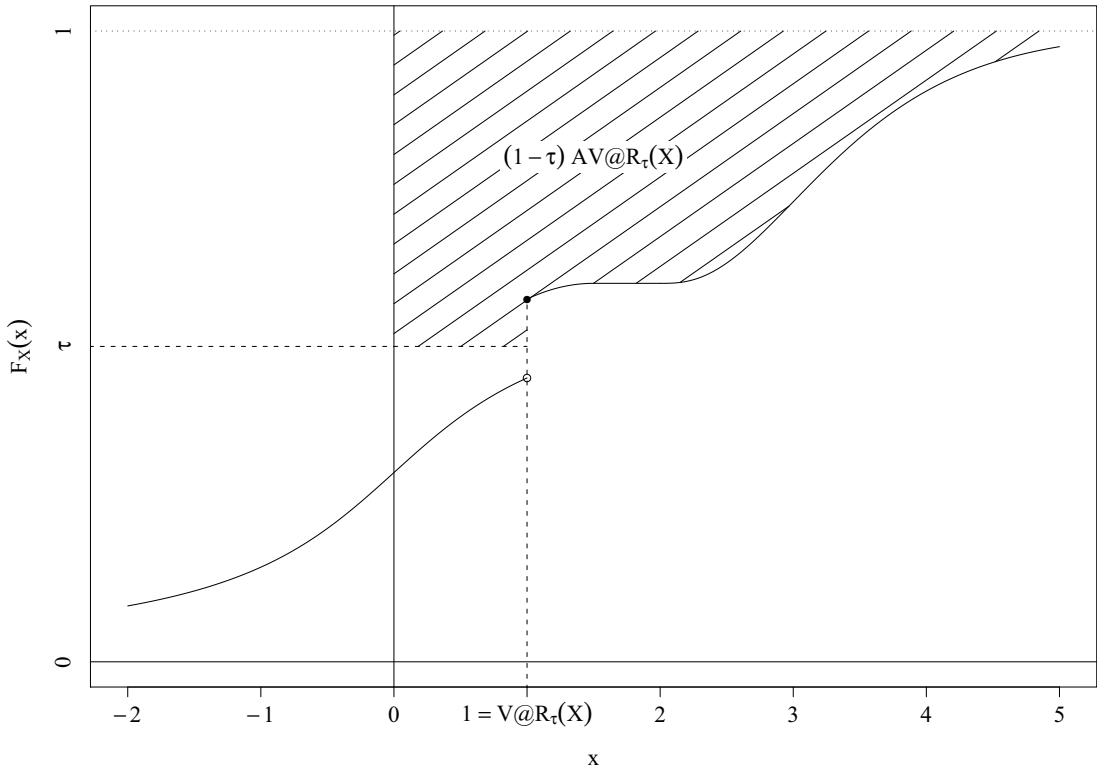


Figure 6.1.: V@R_τ and AV@R_τ.

Theorem 6.2.3. *Let $X \in L^1(\Omega, \mathscr{A}, P)$ and $\tau \in (0,1)$. Then it holds*

$$\text{AV@R}_\tau(X) = \min_{x \in \mathbb{R}} \left\{ x + \frac{1}{1-\tau} E\left[(X-x)^+\right] \right\},$$

where $x^+ := \max\{0, x\}$, $x \in \mathbb{R}$. Moreover,

$$\text{V@R}_\tau(X) = \inf \arg\min_{x \in \mathbb{R}} \left\{ x + \frac{1}{1-\tau} E\left[(X-x)^+\right] \right\}.$$

Proof. See (Rockafellar and Uryasev, 2002). □

So, by Theorem 6.2.3, the computation of the average value-at-risk can be achieved by solving a convex minimization problem rather than by computing an integral of the quantile function. In addition, we obtain the value-at-risk at level τ as the minimal minimizer of this optimization problem.

Coming back to the surveillance task, a policy which minimizes the average value-at-risk at level 0.95 would avoid very devastating threat events because also the highest possible costs come into consideration when using the average value-at-risk. Hence, average value-at-risk could be used as a measure of risk for an application of the threat model.

The average value-at-risk finds further use, e. g., in crop insurance (cf. (Liu et al., 2008)), in operational planning (cf. (Verderame and Floudas, 2010)), in traffic network design and in cancer treatment (cf. (Cromvik and Patriksson, 2010)). In the first application, average value-at-risk is not the criterion itself. Here, average value-at-risk is part of the constraints in corresponding optimization problems. In (Cromvik and Patriksson, 2010), average value-at-risk is used as the objective in a static probabilistic framework. In (Webby et al., 2007), both value-at-risk and average value-at-risk are applied in a case study on the Mekong fishery. They consider four scenarios where they examine how different payment strategies of international aid influence the risk of the fishery. Moreover, value-at-risk and average value-at-risk are used to deal with uncertainty in breeding values of bulls (cf. (Pruzzo et al., 2003)). Of course, the last two articles examine effects in fields that heavily violate animal rights (cf. (Singer, 2002)) and should therefore just be seen as illustrating examples.

7. Average Value-at-Risk Criterion for the Total Discounted Cost – A Non-Standard Objective Criterion for Discrete-Time Markov Decision Processes

In this chapter, we examine a novel objective criterion for MDPs. MDPs are introduced in (Bellman, 1952). We assume that at every time step the decision maker has to pay a random cost depending on the current state and the chosen action. The subsequent costs are discounted and summarized leading to the total discounted cost.

This chapter is structured as follows. Section 7.1 gives an overview on non-standard criteria for MDPs. Section 7.2 provides the main definitions which find use in the following text. In section 7.3, we make some notes concerning the surveillance task which is modelled in continuous time whereas the discrete-time case is examined here. In section 7.4, we formulate the multi-stage optimization problem. In section 7.5, we study the average value-at-risk criterion for the total discounted cost for a finite horizon. This is done in two steps: in the first step, we solve an arising intermediate criterion for a given MDP. In the second step, we derive optimal policies for the average value-at-risk criterion from the solution of the intermediate criterion. This section is concluded with a numerical example. In section 7.6, we treat the infinite-horizon case. Again, this is done by solving an intermediate MDP criterion at first, and then deriving an optimal policy from the solution of the intermediate criterion. In the last section, we discuss the results when the average value-at-risk criterion is applied to the surveillance task.

7.1. Literature Overview on Non-Standard Criteria for Discrete-Time Markov Decision Processes for the Total Discounted Cost

Up to the present, numerous non-standard criteria for MDPs considering the total discounted cost have been studied for the discrete-time case and for the continuous-time case in the literature. Some of these are shortly summarized in the following.

First, (Howard and Matheson, 1972) consider the case of maximizing an exponential utility function with constant risk aversion which leads to a risk-sensitive criterion for discrete-time MDPs. They examine inhomogeneous and homogeneous MDPs with finite and infinite horizon and show that a backward induction algorithm is valid for this problem. They also give a policy improvement algorithm for the homogeneous case which converges to a stationary optimal policy.

Furthermore, (Jaquette, 1973) discusses a criterion called moment-optimality in discrete time for homogeneous MDPs where the moments of the total discounted cost are considered. In this criterion, a policy performs better than another by definition if the signed moments of the first policy are higher than those of the second one with respect to the lexicographical order. This means optimal policies with respect to the expected cost criterion are distinguished by which one has minimal variance. If the variances are also the same, then a policy leading to the maximal third moment of the total discounted cost is better, and so forth. He shows that there exists a stationary optimal policy within the class of deterministic policies in the case of an infinite horizon. By lexicographically ordering the moments of the total discounted cost, some risk aversion comes into play since if the expected value is maximal for some policies, then a policy with minimal variance is chosen. In (Jaquette, 1975), the continuous-time case of the same criterion is considered. The results are similar to the ones from the discrete-time case, i. e., there exists a stationary optimal policy within the class of piecewise constant policies.

Generalizations of the moment-optimality criterion can be found in (Jaquette, 1976; Bouakiz and Sobel, 1992) where the case of optimizing some exponential utility function with a constant risk aversion factor and a constant discount factor is considered. In this case, it turns out that there is no stationary optimal policy within the class of deterministic policies for the infinite-horizon problem in general. But there exist so-called ultimately stationary optimal policies, i. e., the policy is stationary from a certain point in time if the state space is finite. However, a stationary policy can be shown to be optimal in special cases, as for example in (Bouakiz and Sobel, 1992). In (Chung and Sobel, 1987), the same criterion is examined. They provide a backward induction algorithm for the finite-horizon problem, an optimality equation for the optimal value function and approximations for the infinite-horizon problem.

A similar criterion arises as an application in (Porteus, 1975), which also includes maximization of exponential utility. In his framework, exponential utility of the rewards at each time step are considered in contrast to the exponential utility of the total discounted reward. Results are that there exists a stationary optimal policy in the finite-state finite-action case if the risk aversion is constant over time. Furthermore, value iteration can be applied and an optimality equation holds for the optimal value function.

Another criterion is the so-called target level or target value criterion, in which the decision maker wants to find a policy so that the total discounted reward exceeds a given target value with maximal probability. This problem is studied in (Bouakiz and Kebir, 1995; Wu and Lin, 1999). It turns out that a backward induction algorithm holds in the finite-horizon case and that the optimal value function associated with this problem satisfies an optimality equation in the infinite-horizon case and that the optimal value function is its unique solution in a certain class of functions.

The target hitting time criterion is studied in (Boda et al., 2004) which is related to the target level criterion. The objective is to find a policy which minimizes the probability that the total discounted reward does not exceed a given target value at the first $n \in \mathbb{N}$ time steps. Results in (Boda et al., 2004) include a backward induction algorithm and the characterization of optimal policies.

Based on the target value criterion, together with the so-called target percentile criterion the cases of value-at-risk and average value-at-risk are treated in (Boda and Filar, 2006) for the finite horizon, too. They characterize optimal policies for the value-at-risk criterion. Moreover, they show that there are optimal policies which are optimal for the value-at-risk criterion for all confidence levels under strong assumptions. Furthermore, they are able to show that under certain strong conditions there exist time-consistent optimal policies for the average value-at-risk criterion. They do not consider the case of general finite-state finite-action MDPs. Furthermore, the infinite-horizon case is not studied. Their approach is to use the definition of a time-consistent risk measure, whereas we tackle the average value-at-risk criterion my means of an intermediate criterion.

(Schäl, 2004) considers the minimization of the ruin probability of a discrete-time insurance model, where the state space is uncountable, over an infinite horizon. The main results are that the value iteration hold, that there is an optimality equation, and that there is an optimality criterion.

The aforementioned criteria just concern standard MDPs in which the decision maker has to pay one cost or gains one reward per time step respectively. In constraint MDPs, for example, the initial situation is different. In such problems, reward has to be maximized. But together with a single reward, certain costs have to be paid simultaneously. The incurred costs have to satisfy some other conditions like remaining below a certain threshold in expectation. More on this type of problems can be found in (Altman, 1999). Another criterion is studied in (Feinberg and Shwartz, 1994) in which the decision maker yields multiple rewards. Then the decision maker wants to find a policy so that the sum of several total discounted rewards is maximized where different discount factors might be used for each total discounted reward. The results are rather different from the results of the former problems: there might exist some positive ε^* such that for every ε being smaller than ε^* there does not exist a stationary ε-optimal policy. But optimal policies exist under weak conditions.

The preceding criteria are just a small collection of non-standard criteria for MDPs. Further non-standard criteria for MDPs considering the total discounted cost can be found in the survey article (White, 1988).

7.2. Definitions

In this section, we give the main definitions which find use in the further text. Throughout this chapter, let $\Gamma = (S, A, D, W, P, \beta)$ be a *homogeneous finite-state finite-action discrete-time Markov decision process (MDP).* The meanings of the components of Γ are as follows:

- Let S be a non-empty finite *state space.*
- Let A be a non-empty finite *action space.*
- Let $D \subset S \times A$ be the *restriction set,* which satisfies the condition $D(s) := \{a : (s,a) \in D\} \neq \emptyset$, $s \in S$. The set $D(s)$ contains all actions which are admissible in state $s \in S$.
- Let $W \subset \mathbb{R}_{\geq 0}$ be a non-empty finite set, the elements of which should be interpreted as *costs.*
- Let $P = (p^a_{ss'c})_{s,s' \in S, a \in D(s), c \in W}$ be the *transition probabilities.* For s, $s' \in S$, $a \in D(s)$, $c \in W$, the value $p^a_{ss'c}$ is the probability of moving from s to s' resulting in the cost c when action a is taken. Since P contains conditional probabilities, $p^a_{ss'c} \geq 0$ for all s, $s' \in S$, $c \in W$, $a \in D(s)$ and $\sum_{s' \in S, c \in W} p^a_{ss'c} = 1$ for all $(s,a) \in D$.
- Let $\beta \in (0,1)$ be a *discount factor.*

The notation of the transition probabilities $p^a_{ss'c}$ is not the standard definition of the transition probabilities where the costs are fixed for the state and the chosen action. But if the costs are deterministic, i. e., $p^a_{ss'c} = 1$ for some $c \in W$ for all s, $s' \in S$ and $a \in D(s)$, then the model coincides with the standard model. Moreover, the standard notation can be achieved by combining states from S and costs from W to the new state space $S' := S \times W$ and defining the transition probabilities appropriately by $(p')^a_{(s,c)(s',c')} := p^a_{ss'c'}$ for all $c \in W$. The kind of model with random costs as it is used in this text is also considered in (Bertsekas, 2005, 2001; Chung and Sobel, 1987; Wu and Lin, 1999) for instance. Furthermore, (Chung and

Sobel, 1987) give an example why the notation $p^a_{ss'c}$ makes sense in some risk-sensitive models. The example contains an inventory model where the demand is random, and therefore, the reward is random in some states, too. The standard notation comes into play when considering a criterion based on the expectation of the costs. Then one defines transition probabilities $p^a_{ss'} := \sum_{c\in W} p^a_{ss'c}$, s, $s' \in S$, $a \in D(s)$, and expected costs $c(s,a) := \sum_{c\in W, s'\in S} p^a_{ss'c}\, c$, $(s,a) \in D$, leading to the standard MDP definition. But these definitions would not meet the demands of a risk-sensitive criterion since the distribution of the one-step costs plays a crucial role. Therefore, we use the notation $p^a_{ss'c}$. Homogeneity of the MDP means that the transition probabilities are not time-dependent.

Next, we model the basis of decision-making. We assume that the decision maker has access to the whole past at any point in time and the current state but that she cannot look into the future. This basis is given by a history which is defined next.

Definition 7.2.1. Define for $k \in \mathbb{N}_0$ the *sets of histories (up to time step k)* recursively by

$$H_0 := S,$$
$$H_{k+1} := H_k \times A \times W \times S.$$

An element of the set H_k is called a *history (up to time step k).* For every $k \in \mathbb{N}$, a history $h_k = (x_0, a_0, c_0, x_1, \ldots, a_{k-1}, c_{k-1}, x_k) \in H_k$ is called *admissible* if $a_l \in D(x_l)$ for all $0 \le l \le k-1$.

A history $h_k = (x_0, a_0, c_0, x_1, \ldots, a_{k-1}, c_{k-1}, x_k) \in H_k$, $k = 0, 1, \ldots$, is admissible if all actions a_l are in the restriction set $D(x_l)$, $0 \le l \le k-1$, of the current state. Based on the knowledge of the past and the current state, the decision maker is able to choose the next action. We assume that the decision maker can randomly choose between admissible actions according to a selected probability distribution on the admissible actions. Therefore, an admissible policy is defined as a randomized history-dependent policy.

Definition 7.2.2. A *randomized history-dependent policy* $\pi = (\pi_k)_{k\in\mathbb{N}_0}$ is a sequence of measurable mappings $\pi_k : H_k \times A \to [0,1]$ such that

$$\sum_{a\in D(x_k)} \pi_k(h_k, a) = 1 \quad \text{and} \quad \pi_k(h_k, a) = 0 \quad (a \notin D(x_k))$$

for all $k \in \mathbb{N}_0$ and for all admissible $h_k = (x_0, a_0, c_0, x_1, \ldots, a_{k-1}, c_{k-1}, x_k) \in H_k$. We also write $\pi_k(a|h_k) := \pi_k(h_k, a)$. Further, let Π be the *set of all randomized history-dependent policies.*

We show that for every randomized history-dependent policy $\pi \in \Pi$ there is a probability space $(\Omega, \mathscr{A}, P^\pi)$ so that the parameters of the MDP match their interpretations. The construction of these probability spaces is according to (Puterman, 2005), section 2.1.6. Let

$$\Omega := \bigtimes_{k=0}^{\infty} (S \times A \times W) \quad \text{and} \quad \mathscr{A} := \bigotimes_{k=0}^{\infty} (\mathscr{P}(S) \times \mathscr{P}(A) \times \mathscr{P}(W)), \quad \pi \in \Pi,$$

where $\mathscr{P}(M)$ denotes the power set of the set M. Let $\omega \in \Omega$ be of the form

$$\omega = (x_0, a_0, c_0, x_1, a_1, c_1, \ldots).$$

Define the projections $X_k(\omega) := x_k$, $A_k(\omega) := a_k$, $C_k(\omega) := c_k$, $k \in \mathbb{N}_0$. So, $(X_k)_{k\in\mathbb{N}_0}$, $(A_k)_{k\in\mathbb{N}_0}$ and $(C_k)_{k\in\mathbb{N}_0}$ are the state, action and costs processes respectively. Let P_0 be the initial probability distribution, which we assume to be concentrated on a single state in the whole chapter so that $P_0(\{x_0\}) = 1$ for some fixed $x_0 \in S$. A policy $\pi \in \Pi$ induces a probability measure P^π on $\mathscr{A}$ by setting

$$\begin{aligned}
&P^\pi(X_0 = x_0) = P_0(\{x_0\}),\\
&P^\pi(A_k = a_k \,|\, X_0 = x_0, A_0 = a_0, C_0 = c_0, \ldots, X_k = x_k) = \pi_k(a_k|x_0, a_0, c_0, \ldots, x_k),\\
&P^\pi(C_k = c_k, X_{k+1} = x_{k+1} \,|\, X_0 = x_0, A_0 = a_0, c_0 = c_0, \ldots, X_k = x_k, A_k = a_k) = p^{a_k}_{x_k x_{k+1} c_k},\\
&\qquad\qquad x_0, \ldots, x_{k+1} \in S,\ a_0, \ldots, a_k \in A,\ c_0, \ldots, c_k \in W,\ k = 0, 1, \ldots,
\end{aligned} \tag{7.1}$$

from which

$$\begin{aligned}
&P^\pi(X_0 = x_0, A_0 = a_0, c_0 = c_0, \ldots, C_k = c_k, X_{k+1} = x_{k+1})\\
&\quad = P_1(x_0)\, \pi_0(a_0|x_0)\, p^{a_0}_{x_0 x_1 c_0} \cdot \ldots \cdot \pi_k(a_k|x_0, a_0, c_0, \ldots, x_k)\, p^{a_k}_{x_k x_{k+1} c_k},
\end{aligned}$$

$$x_0, \dots, x_{k+1} \in S,\ a_0, \dots, a_k \in A,\ c_0, \dots, c_k \in W,\ k = 0, 1, \dots.$$

By this equation, the finite-dimensional distributions of the first components are uniquely determined by a theorem of Ionescu-Tulcea (cf. (Bertsekas and Shreve, 1978), Proposition 7.28), furthermore, P^π is the unique probability measure on $(\Omega, \mathscr{A})$ such that (7.1) holds. This completes the construction of the probability space $(\Omega, \mathscr{A}, P^\pi)$.

The set of randomized history-dependent policies is the largest class of policies we consider in this section. In the case of the expected total discounted cost criterion, it is known that there exist optimal policies which have much more structure. The next definition gives a variety of policies, which we come across when establishing the following results.

Definition 7.2.3. Let $\pi \in \Pi$. For $k \in \mathbb{N}_0$, let $h_k \in H_k$ be of the form $h_k = (x_0, a_0, c_0, x_1, \dots, a_{k-1}, c_{k-1}, x_k)$.

- The policy π is a *deterministic history-dependent policy,* denoted by $\pi \in \Pi^d$, if for all $k \in \mathbb{N}_0$ and $h_k \in H_k$ there exists some $a \in D(x_k)$ such that $\pi_k(a|h_k) = 1$. In this case, we also write $\pi_k(h_k) = a$. A deterministic history-dependent policy chooses a certain action with probability one depending on the history.

- The policy π is a *randomized Markovian policy,* denoted by $\pi \in \Pi_m$, if for every $k \in \mathbb{N}_0$ it holds $\pi_k(a|h_k) = \pi_k(a|x'_0, a'_0, c'_0, x'_1, \dots, a'_{k-1}, c'_{k-1}, x_k) =: \pi_k(a|x_k)$ for all $(x'_l, a'_l) \in D$, $c'_l \in W$, $0 \le l \le k-1$. A randomized Markovian policy randomly chooses an action only depending on the current state and time step.

- The policy π is a *deterministic Markovian policy,* denoted by $\pi \in \Pi_m^d$, if $\pi \in \Pi^d \cap \Pi_m$. Then we write $\pi_k(x_k) = \pi_k(h_k)$. A deterministic policy chooses a certain action with probability one only depending on the current state and time step.

- The policy π is a *randomized stationary policy,* denoted by $\pi \in \Pi_s$, if $\pi \in \Pi_m$ and $\pi_0 = \pi_1 = \pi_2 = \dots$. Then we identify the policy π with a component π_k. A randomized stationary policy chooses randomly an action only depending on the current state and independent of the time step.

- The policy π is a *deterministic stationary policy,* denoted by $\pi \in \Pi_s^d$, if $\pi \in \Pi^d \cap \Pi_s$. If $\pi_k(a|x_k) = 1$ for some $a \in D(x_k)$, then we also write $\pi(x_k) = a$. A deterministic stationary policy chooses a certain action with probability one only depending on the current state and independent of the time step.

- A mapping $\mu : S \to A$ with $\mu(s) \in D(s)$ for all $s \in S$ is called a *decision rule.* Let $F := \{\mu \,|\, \mu : S \to A,\ \mu(s) \in D(s)\ (s \in S)\}$ be the *set of decision rules.* For $\mu \in F$, the policy $\mu^\infty := (\mu, \mu, \dots)$ is an element of Π_s^d, and we identify μ^∞ with μ.

We have the following implications:

$$\Pi_s^d \subset \Pi_s \subset \Pi_m \subset \Pi \quad \text{and} \quad \Pi_s^d \subset \Pi_m^d \subset \Pi^d \subset \Pi.$$

In this chapter, discounted costs are considered. The *total discounted cost up to time step* $n \in \mathbb{N}_0$ and the *total discounted cost over an infinite horizon* are defined by

$$C^n := \sum_{k=0}^{n} \beta^k C_k \quad \text{and} \quad C^\infty := \sum_{k=0}^{\infty} \beta^k C_k,$$

respectively. The discounting with factor $\beta \in (0,1)$ leads to a decrease in the weighting of the costs which are incurred at a later time. Under every policy $\pi \in \Pi$ and every initial state $x_0 \in S$, the total discounted costs C^n and C^∞ are random variables with certain conditional distributions given by their conditional cumulative distribution functions $P^\pi(C^n \le \xi \,|\, X_0 = x_0)$ and $P^\pi(C^\infty \le \xi \,|\, X_0 = x_0)$, $\xi \in \mathbb{R}$, respectively. Since W is assumed to be finite, we define the upper bound of W by $U := \max W$. Hence, for every $\pi \in \Pi$ we have

$$0 \le C^n \le \frac{1-\beta^{n+1}}{1-\beta} U \quad P^\pi\text{-a. s.}, \quad n \in \mathbb{N}_0, \qquad \text{and}$$
$$0 \le C^\infty \le \frac{U}{1-\beta} \quad P^\pi\text{-a. s.}.$$

We make use of these relationships in the further text.

7.3. Continuous- and Discrete-Time Markov Decision Processes

The model of the dynamics of threat of an infrastructure is defined in terms of a CMDP (cf. chapter 3). But we can reformulate the model as a discrete-time MDP by using a uniformization technique as it is proposed in (Serfozo, 1979). Assume we have a finite state space and a finite action space. Let $\tilde{\lambda}(s,a) \coloneqq \tilde{\lambda} \coloneqq \max_{(s',a')\in D} \lambda(s',a')$, $(s,a) \in D$, be the transition rate for the uniformized model. In the uniformized model, the transition probabilities $(\tilde{p}^a_{ss'})_{s,s'\in S, a\in D(s)}$ and the costs $(\tilde{c}(s,a))_{(s,a)\in D}$ are defined by

$$\tilde{p}^a_{ss'} \coloneqq \begin{cases} 1 - \frac{\lambda(s,a)\,(1-p^a_{ss})}{\tilde{\lambda}}, & \text{if } s' = s \\ \frac{\lambda(s,a)\,p^a_{ss'}}{\tilde{\lambda}}, & \text{if } s' \neq s \end{cases},$$

$$\tilde{c}(s,a) \coloneqq \frac{\alpha + \lambda(s,a)}{\alpha + \tilde{\lambda}}\, c(s,a)$$

for s, $s' \in S$ and $a \in D(s)$. The discount rate for the uniformized model is again α. As it can be seen from the uniformization procedure, the original model is varied such that for every state and every action the transition rate is the same. In the uniformized model, a number of additional "fictitious" state transitions from one state to the same occur. In order to conserve the probabilistic behaviour under deterministic stationary policies, the transition probabilities and costs have to be changed appropriately. (Serfozo, 1979) shows that the expectations of the total discounted cost of every deterministic stationary policy coincide under the continuous-time and under the uniformized model. (Puterman, 2005) generalizes this fact to randomized stationary policies in Proposition 11.5.1. In order to fit the uniformized model to the non-standard MDP formulation used here, we have to define the transition probabilities $\hat{p}^a_{ss'\tilde{c}(s,a)} = \tilde{p}^a_{ss'}$, $(s,a) \in D$.

Therefore, a deterministic stationary optimal policy for the continuous-time problem with expected total discounted cost criterion can be derived by solving the uniformized version of the problem and taking a deterministic stationary optimal policy of the uniformized model as a solution for the continuous-time problem. This works because it is known that in both cases there exists a deterministic stationary optimal policy. Moreover, the optimality equation of the uniformized model matches the optimality of a discrete-time MDP with discount factor $\tilde{\lambda}/(\alpha+\tilde{\lambda})$, and hence methods for discrete-time MDPs can be used to obtain a deterministic stationary optimal policy.

But this does not justify that we can use the uniformization technique without any further consideration in order to obtain an optimal policy for the average value-at-risk criterion. If there is no deterministic stationary optimal policy for the uniformized model, first, it might not be clear how to convert the optimal policy for the uniformized model into a policy for the original continuous-time model. Second, if such a conversion is feasible, then it is not clear a priori that the converted policy is optimal for the continuous-time model.

Although the effect of the uniformization technique on the preceding results is not known and is not studied in this text, we assume that the data of the surveillance task are given as a discrete-time MDP and that the results of this chapter are considerably good when applied to the surveillance task.

7.4. The Average Value-at-Risk Criterion

The average value-at-risk is besides expectation, moments, variance and quantiles another characteristic of a random variable. Given the confidence level $\tau \in (0,1)$, the average value-at-risk at level τ is the average of the $(1-\tau)\cdot 100\,\%$ highest costs. The definition of the average value-at-risk has already been given in Definition 6.2.2.

For $\pi \in \Pi$, $x_0 \in S$, $n \in \mathbb{N}_0$ and $\tau \in (0,1)$, we write $\mathrm{AV@R}^{\pi}_{\tau}(C^n \,|\, X_0 = x_0)$ for the average value-at-risk of C^n at level τ under policy π if the initial state is x_0 and $\mathrm{AV@R}^{\pi}_{\tau}(C^{\infty} \,|\, X_0 = x_0)$ in the infinite-horizon case. In the context of the defined MDP, the average value-at-risk problem is the following for a finite horizon $n \in \mathbb{N}_0$: find a policy $\pi^* \in \Pi$, if it exists, such that

$$\mathrm{AV@R}^{\pi^*}_{\tau}(C^n \,|\, X_0 = x_0) = \inf_{\pi\in\Pi} \mathrm{AV@R}^{\pi}_{\tau}(C^n \,|\, X_0 = x_0) =: \mathrm{AV@R}^{*}_{\tau}(C^n \,|\, X_0 = x_0)\,. \tag{7.2}$$

For the infinite horizon, the problem we want to solve is: find a policy $\pi^* \in \Pi$, if it exists, such that

$$\mathrm{AV@R}^{\pi^*}_{\tau}(C^{\infty} \,|\, X_0 = x_0) = \inf_{\pi\in\Pi} \mathrm{AV@R}^{\pi}_{\tau}(C^{\infty} \,|\, X_0 = x_0) =: \mathrm{AV@R}^{*}_{\tau}(C^{\infty} \,|\, X_0 = x_0)\,. \tag{7.3}$$

Definition 7.4.1. A policy $\pi^* \in \Pi$ satisfying (7.2) for all $x_0 \in S$ is *optimal with respect to the average value-at-risk criterion at confidence level τ for the discounted cost for the finite horizon n.* A policy $\pi^* \in \Pi$ satisfying (7.3) for all $x_0 \in S$ is *optimal with respect to the average value-at-risk criterion at confidence level τ for the discounted cost for the infinite horizon.*

From now on, let the initial state $x_0 \in S$ and level $\tau \in (0,1)$ be fixed. By Theorem 6.2.3, we can rewrite the right-hand sides of the problems (7.2) and (7.3) as

$$\begin{aligned}\inf_{\pi\in\Pi} \mathrm{AV@R}_\tau^\pi(C^n \,|\, X_0 = x_0) &= \inf_{\pi\in\Pi} \min_{\xi\in\mathbb{R}} \left\{ \xi + \frac{1}{1-\tau} E^\pi \left[(C^n - \xi)^+ \,\middle|\, X_0 = x_0 \right] \right\} \\ &\overset{\text{Lemma A.4}}{=} \inf_{\xi\in\mathbb{R}} \inf_{\pi\in\Pi} \left\{ \xi + \frac{1}{1-\tau} E^\pi \left[(C^n - \xi)^+ \,\middle|\, X_0 = x_0 \right] \right\} \\ &= \inf_{\xi\in\mathbb{R}} \left\{ \xi + \frac{1}{1-\tau} \inf_{\pi\in\Pi} E^\pi \left[(C^n - \xi)^+ \,\middle|\, X_0 = x_0 \right] \right\} \end{aligned} \tag{7.4}$$

in the finite-horizon case and in the case of an infinite horizon as

$$\begin{aligned}\inf_{\pi\in\Pi} \mathrm{AV@R}_\tau^\pi(C^\infty \,|\, X_0 = x_0) &= \inf_{\pi\in\Pi} \min_{\xi\in\mathbb{R}} \left\{ \xi + \frac{1}{1-\tau} E^\pi \left[(C^\infty - \xi)^+ \,\middle|\, X_0 = x_0 \right] \right\} \\ &\overset{\text{Lemma A.4}}{=} \inf_{\xi\in\mathbb{R}} \inf_{\pi\in\Pi} \left\{ \xi + \frac{1}{1-\tau} E^\pi \left[(C^\infty - \xi)^+ \,\middle|\, X_0 = x_0 \right] \right\} \\ &= \inf_{\xi\in\mathbb{R}} \left\{ \xi + \frac{1}{1-\tau} \inf_{\pi\in\Pi} E^\pi \left[(C^\infty - \xi)^+ \,\middle|\, X_0 = x_0 \right] \right\}, \end{aligned} \tag{7.5}$$

simply by interchanging the minimization procedures. Hence, in order to obtain solutions for the original problems, i. e., the average value-at-risk problems, we can solve the inner problems in (7.4) and (7.5) for all $\xi \in \mathbb{R}$ in advance. Afterwards, we pick some $\xi^* \in \mathbb{R}$ such that it minimizes equation (7.4) for the finite horizon and (7.5) for the infinite horizon, respectively, which exists in both cases. This procedure leads to the optimal values of $\mathrm{AV@R}_\tau^*(C^n \,|\, X_0 = x_0)$ and $\mathrm{AV@R}_\tau^*(C^\infty \,|\, X_0 = x_0)$. Furthermore, we are able to construct an optimal policy for the average value-at-risk criterion by executing these steps.

In the next section, the finite-horizon problem is examined, followed by the section which treats the infinite-horizon case.

7.5. The Finite Horizon

The case of a finite horizon is treated in (Boda and Filar, 2006). Their result is that under certain strong conditions there exists a time-consistent optimal policy for the average value-at-risk criterion in the finite-horizon case. But the structure of such optimal policies remains unclear and hard to compute. Their approach is based on the cumulative distribution function of the total discounted cost which leads to the time-consistent risk measure which they call target percentile risk measure.

The main result of this section is that there is a deterministic, in general not Markovian, optimal policy for the average value-at-risk criterion. In contrast to (Boda and Filar, 2006), the results are derived by solving intermediate problems for MDPs first for which we can prove the existence of deterministic optimal policies. The details and the derivation follow.

Although we have a look at a finite horizon only, we consider a general infinite-horizon policy $\pi \in \Pi$ as defined in Definition 7.2.2 where actions are chosen for every $n \in \mathbb{N}$. But for a horizon consisting of $n \in \mathbb{N}_0$ time steps, we are just interested in the first n entries of π. The tail of policies concerning the n-horizon problem does not play any role in our consideration and, hence, can be chosen arbitrarily. But if the n first components of π in an n-horizon problem satisfy any of the conditions given in Definition 7.2.3, we also say that π is of the respective type if its first n components satisfy the respective definition.

7.5.1. An Intermediate Criterion

From (7.4), it can be seen that minimizing average value-at-risk can be done by two optimization steps. First, we consider the inner optimization problem. Fix $n \in \mathbb{N}_0$ and $\xi \in \mathbb{R}$ and define the *n-horizon value function* for an arbitrary policy $\pi \in \Pi$ by

$$v_n^\pi(x_0, \xi) := E^\pi \left[(C^n - \xi)^+ \,|\, X_0 = x_0 \right], \quad x_0 \in S.$$

Then the inner problem of (7.4) is:

$$\text{Minimize } v_n^\pi(x_0, \xi) \text{ over all } \pi \in \Pi. \tag{7.6}$$

Note that (7.6) is independent of level τ. Minimizing $v_n^\pi(x_0,\xi)$ is another criterion for MDPs in its own right. Considering this criterion, only the costs exceeding the threshold value ξ are penalized by the amount of excess. If C^n remains below ξ, the decision maker is not penalized. For $\xi \leq 0$, problem (7.6) is equivalent to the expected total discounted cost criterion since C^n is considered to be non-negative. Hence, there exists a stationary optimal decision rule which solves (7.6) by (Puterman, 2005), Theorem 6.2.10, for example. But for $\xi > 0$, problem (7.6) differs from the expected total discounted cost criterion. The decision maker considering the intermediate criterion has to take into account that the costs up to ξ are not penalized since the contribution to $v^\pi(x_0,\xi)$ is zero. But if $\xi > U/(1-\beta)$, then $(C^n-\xi)^+ = 0$ P^π-a. s. under every policy $\pi \in \Pi$, and hence every policy is optimal. A trivial observation is that the value function satisfies $v^\pi(x_0,\xi) \geq 0$ for all $x_0 \in S$, $\pi \in \Pi$ and for all $\xi \in \mathbb{R}$.

Now, define for $\pi \in \Pi$, $(x_0,a_0) \in D$, and $c_0 \in W$ a policy ${}^1\pi^{(x_0,a_0,c_0)} \in \Pi$ by

$$ {}^1\pi_k^{(x_0,a_0,c_0)}(a|h_k) := \pi_{k+1}(a|(x_0,a_0,c_0,h_k)), \quad a \in A,\, h_k \in H_k,\, k \in \mathbb{N}_0. $$

The policy ${}^1\pi^{(x_0,a_0,c_0)}$ is called the *cut-head policy of* π (cf. (Wu and Lin, 1999)). The cut-head policy based on $\pi \in \Pi$ and $(x_0,a_0,c_0) \in S \times A \times W$ acts according to π shifted one time step forward in time, when (x_0,a_0,c_0) is observed at the 0th time step. With this definition, we prove the following essential lemma.

Lemma 7.5.1. *Let $\pi \in \Pi$. Then for every $n \in \mathbb{N}$, it holds*

$$ v_n^\pi(s,\xi) = \beta \sum_{a_0 \in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0}\, v_{n-1}^{{}^1\pi^{(s,a_0,c_0)}}\left(s', \frac{\xi - c_0}{\beta}\right), \quad s \in S,\, \xi \in \mathbb{R}. $$

Proof. Let $\pi \in \Pi$, $n \in \mathbb{N}$, $s \in S$ and $\xi \in \mathbb{R}$. Then computation yields by positive homogeneity of $(\cdot)^+$ (cf. Lemma A.2) and the definition of ${}^1\pi^{(s,a_0,c_0)}$

$$ \begin{aligned} v_n^\pi(s,\xi) &= E^\pi\left[(C_0 + \beta C_1 + \ldots + \beta^n C_n - \xi)^+ \,\middle|\, X_0 = s\right] \\ &= \sum_{a_0 \in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} E^\pi\left[(c_0 + \beta C_1 + \ldots + \beta^n C_n - \xi)^+ \,\middle|\, X_0 = s, A_0 = a_0, C_0 = c_0, X_1 = s'\right] \\ &= \beta \sum_{a_0 \in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} E^\pi\left[\left(C_1 + \ldots + \beta^{n-1} C_n - \frac{\xi - c_0}{\beta}\right)^+ \,\middle|\, X_0 = s, A_0 = a_0, C_0 = c_0, X_1 = s'\right] \\ &= \beta \sum_{a_0 \in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} E^{{}^1\pi^{(s,a_0,c_0)}}\left[\left(C_0 + \ldots + \beta^{n-1} C_{n-1} - \frac{\xi - c_0}{\beta}\right)^+ \,\middle|\, X_0 = s'\right] \\ &= \beta \sum_{a_0 \in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0}\, v_{n-1}^{{}^1\pi^{(s,a_0,c_0)}}\left(s', \frac{\xi - c_0}{\beta}\right), \end{aligned} $$

which is the assertion. □

We denote Lemma 7.5.1 in short notation by

$$ v_n^\pi = T_{\pi_0} v_{n-1}^{{}^1\pi}, \quad \pi \in \Pi,\, n \in \mathbb{N}. $$

The 0-Horizon Case

In this section, we restrict ourselves to the case $n = 0$ to give a first idea of the general finite-horizon problem. That is, the decision maker has to make exactly one decision at time step 0. The system starts in $x_0 \in S$. According to the decision maker, an action a_0 is chosen, possibly by a random mechanism, a cost c_0 occurs, and the system ends up in a final state s'. The following proposition characterizes the structure of an optimal policy of problem (7.6).

Proposition 7.5.2. *Let $x_0 \in S$, $\xi \in \mathbb{R}$ and $n = 0$. Then a policy $\pi^* \in \Pi$ is optimal for problem* (7.6) *if and only if*

$$ \pi_0^*(a|x_0) = 0 \quad \textit{for every} \quad a \notin \operatorname*{arg\,min}_{a_0 \in D(x_0)} \left\{ \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{x_0 s' c_0}^{a_0} (c_0 - \xi)^+ \right\}. $$

Proof. The assertion follows immediately from

$$v_0^\pi(x_0,\xi) = E^\pi\left[\left(C^0-\xi\right)^+ \,\middle|\, X_0 = x_0\right] = \sum_{a_0\in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0\in W,\\ s'\in S}} p^{a_0}_{x_0 s' c_0} (c_0-\xi)^+ \geq \min_{a_0\in D(x_0)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^{a_0}_{x_0 s' c_0} (c_0-\xi)^+$$

for every $\pi\in\Pi$ and $x_0\in S$ since $D(x_0)$ is finite and since equality holds for π_0^* of the proposed form. □

From Proposition 7.5.2, it already follows that there exists a deterministic optimal policy for problem (7.6) since a policy $\pi_\xi^*\in\Pi_s^d$ depending on the threshold $\xi\in\mathbb{R}$ is given by

$$\left(\pi_\xi^*\right)_0(x_0) = a, \quad a\in \operatorname*{arg\,min}_{a_0\in D(x_0)} \left\{ \sum_{\substack{c_0\in W,\\ s'\in S}} p^{a_0}_{x_0 s' c_0} (c_0-\xi)^+ \right\} \text{ arbitrary,}$$

$$\left(\pi_\xi^*\right)_k(x_0) = a_k, \quad a_k\in D(x_0) \text{ arbitrary, } k\geq 1,$$

is optimal.

The General Finite-Horizon Case

In this section, we want to solve problem (7.6) for an arbitrary finite horizon of length $n\in\mathbb{N}$. For $n\in\mathbb{N}_0$, define the *optimal value function* v_n^* by

$$v_n^*(s,\xi) := \inf_{\pi\in\Pi} v_n^\pi(s,\xi), \quad s\in S,\ \xi\in\mathbb{R}.$$

For the proofs, we need some more definitions. Let $v: S\times\mathbb{R}\to\mathbb{R}$ and $\mu\in F$. Then we define one-step operators by

$$T_\mu v(s,\xi) := \beta \sum_{\substack{c_0\in W,\\ s'\in S}} p^{\mu(s)}_{ss'c_0}\, v\left(s', \frac{\xi-c_0}{\beta}\right), \quad s\in S,\ \xi\in\mathbb{R},$$

$$Tv(s,\xi) := \beta \min_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^{a}_{ss'c_0}\, v\left(s', \frac{\xi-c_0}{\beta}\right), \quad s\in S,\ \xi\in\mathbb{R}.$$

Note that a minimizing action a in the definition of T may depend on both s and ξ.

Lemma 7.5.3. *The operators T_μ and T are isotone, i. e., for u, $v: S\times\mathbb{R}\to\mathbb{R}$ with $u\leq v$, it holds $T_\mu u\leq T_\mu v$ and $Tu\leq Tv$.*

Proof. The assertion follows from the definitions of T_μ and T and from the non-negativity of β and $p^a_{ss'c_0}$ for every s, $s'\in S$, $a\in D(s)$ and $c_0\in W$. □

Next, we prove a backward induction procedure for the optimal value functions. This is the essential proposition of this section. Moreover, optimal policies for problem (7.6) are constructed in the course of the proof.

Proposition 7.5.4. *With $v_{-1}^*(s,\xi) := \xi^- := -\min\{0,\xi\}$, $s\in S$, $\xi\in\mathbb{R}$, it holds*

$$v_n^* = Tv_{n-1}^*, \quad n\in\mathbb{N}_0.$$

Moreover, for every $s\in S$ and $\xi\in\mathbb{R}$, there exists a deterministic optimal policy $\pi_\xi^{n,*}\in\Pi^d$ for the n-horizon problem (7.6)*, $n\in\mathbb{N}_0$.*

Proof. We prove the proposition by induction. At first, for $\xi\in\mathbb{R}$ it holds $\xi^+ = (-\xi)^-$. Fix $\xi\in\mathbb{R}$. For $n=0$, define a deterministic optimal policy $\pi_\xi^{0,*}$ according to the preceding section. Then by positive homogeneity of $(\cdot)^+$

$$\begin{aligned} v_0^*(s,\xi) &= \sum_{\substack{c_0\in W,\\ s'\in S}} p^{\left(\pi_\xi^{0,*}\right)_0(s)}_{ss'c_0} \cdot (c_0-\xi)^+ = \min_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{ss'c_0} (c_0-\xi)^+ = \beta \min_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{ss'c_0} \left(\frac{\xi-c_0}{\beta}\right)^- \\ &= Tv_{-1}^*(s,\xi), \end{aligned}$$

$s \in S$, by the definitions of $\pi_\xi^{0,*}$ and T. Hence, the assertion holds true for $n = 0$. Now, let $v_n^* = Tv_{n-1}^*$ be true for some $n \in \mathbb{N}_0$. For every $\xi \in \mathbb{R}$, let $\pi_\xi^{n,*} \in \Pi^d$ with $v_n^{\pi_\xi^{n,*}}(s,\xi) = v_n^*(s,\xi)$ for all $s \in S$ as part of the induction hypothesis (IH). Then for every $s \in S$ and $\xi \in \mathbb{R}$

$$
\begin{aligned}
v_{n+1}^*(s,\xi) &= \inf_{\pi\in\Pi} E^\pi\left[\left(C^{n+1}-\xi\right)^+ \,\middle|\, X_0 = s\right] \\
&\overset{\text{Lemma 7.5.1}}{=} \beta \inf_{\pi\in\Pi} \sum_{a_0\in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0}\, E^{{}^1\pi^{(s,a_0,c_0)}}\left[\left(C_0+\beta C_1+\ldots+\beta^n C_n - \frac{\xi-c_0}{\beta}\right)^+ \,\middle|\, X_0 = s'\right] \\
&\geq \beta \inf_{\pi\in\Pi} \sum_{a_0\in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0}\, v_n^*\left(s', \frac{\xi-c_0}{\beta}\right) = \beta \min_{a_0\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0}\, v_n^*\left(s', \frac{\xi-c_0}{\beta}\right) = Tv_n^*(s,\xi). \qquad (7.7)
\end{aligned}
$$

To see that the reverse inequality holds true, we define for every $\xi \in \mathbb{R}$ a policy $\pi_\xi^{n+1,*} \in \Pi^d$ for which we can show that it is optimal for the $(n+1)$-horizon problem. To this end, let

$$
\begin{aligned}
&\left(\pi_\xi^{n+1,*}\right)_0(s) \in \operatorname*{arg\,min}_{a_0\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0}\, v_n^*\left(s', \frac{\xi-c_0}{\beta}\right) \text{ arbitrary,} \\
&\left(\pi_\xi^{n+1,*}\right)_1(c_0,x_1) \coloneqq \left(\pi_\xi^{n+1,*}\right)_1(s,a_0,c_0,x_1) \coloneqq \left(\pi_{\frac{\xi-c_0}{\beta}}^{n,*}\right)_0(x_1), \\
&\left(\pi_\xi^{n+1,*}\right)_k(c_0,\ldots,c_{k-1},x_k) \coloneqq \left(\pi_\xi^{n+1,*}\right)_k(s,a_0,c_0,x_1,\ldots,a_{k-1},c_{k-1},x_k) \coloneqq \left(\pi_{\frac{\xi-c_0}{\beta}}^{n,*}\right)_{k-1}(c_1,\ldots,c_{k-1},x_k), \\
&\qquad\qquad s, x_1, \ldots, x_k \in S,\ a_0, \ldots, a_{k-1} \in A,\ c_0, \ldots, c_k \in W,\ k = 2, \ldots, n+1,
\end{aligned} \qquad (7.8)
$$

and $\left(\pi_\xi^{n+1,*}\right)_k$ arbitrary for $k \geq n+2$. For $\pi_\xi^{n+1,*}$ and $s, s' \in S$, $a \in D(s)$ and $c, c_0, \ldots, c_{k-1} \in W$, the respective cut-head policies are given by

$$
\begin{aligned}
&{}^1\left(\pi_\xi^{n+1,*}\right)_0^{(s,a,c)}(s') = \left(\pi_\xi^{n+1,*}\right)_1(c,s') = \left(\pi_{\frac{\xi-c}{\beta}}^{n,*}\right)_0(s'), \\
&{}^1\left(\pi_\xi^{n+1,*}\right)_k^{(s,a,c)}(c_0,\ldots,c_{k-1},s') = \left(\pi_\xi^{n+1,*}\right)_{k+1}(c,c_0,\ldots,c_{k-1},s') = \left(\pi_{\frac{\xi-c}{\beta}}^{n,*}\right)_k(c_0,\ldots,c_{k-1},s'), \\
&\qquad\qquad k = 1, \ldots, n+1.
\end{aligned}
$$

So, ${}^1\left(\pi_\xi^{n+1,*}\right)_k^{(s,a,c)} = \left(\pi_{\frac{\xi-c}{\beta}}^{n,*}\right)_k$, $(s,a) \in D$, $c \in W$, $k = 0, \ldots, n+1$. With the abbreviation $a_0(s) \coloneqq \left(\pi_\xi^{n+1,*}\right)_0(s)$, $s \in S$, it follows

$$
\begin{aligned}
v_{n+1}^*(s,\xi) &\leq v_{n+1}^{\pi_\xi^{n+1,*}}(s,\xi) \overset{\text{Lemma 7.5.1}}{=} T_{\left(\pi_\xi^{n+1,*}\right)_0} {}^1v_n^{\pi_\xi^{n+1,*}}(s,\xi) \\
&= \beta \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0(s)}\, E^{{}^1\left(\pi_\xi^{n+1,*}\right)^{(s,a_0(s),c_0)}}\left[\left(C_0+\beta C_1+\ldots+\beta^n C_n - \frac{\xi-c_0}{\beta}\right)^+ \,\middle|\, X_0 = s'\right] \\
&= \beta \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0(s)}\, E^{\pi_{\frac{\xi-c_0}{\beta}}^{n,*}}\left[\left(C_0+\beta C_1+\ldots+\beta^n C_n - \frac{\xi-c_0}{\beta}\right)^+ \,\middle|\, X_0 = s'\right] \\
&\overset{\text{(IH)}}{=} \beta \sum_{\substack{c_0\in W,\\ s'\in S}} p_{ss'c_0}^{a_0(s)}\, v_n^*\left(s', \frac{\xi-c_0}{\beta}\right) = Tv_n^*(s,\xi) \overset{(7.7)}{\leq} v_{n+1}^*(s,\xi).
\end{aligned}
$$

Hence, equality holds in the previous inequality chain. Thus, $\pi_\xi^{n+1,*}$ is an optimal policy for the $(n+1)$-horizon problem, which concludes the proof. □

Remark 7.5.5. 1. The operator T defines an MDP with uncountable Borel state space $S \times \mathbb{R}$. The system jumps from (s,ξ) to $(s',(\xi-c)/\beta)$ with probability $p^a_{ss'c}$ if action $a \in D(s)$ is chosen. The incurred cost is zero for all states and for all actions and the discount factor is β. Theory on discrete-time Markov decision processes with uncountable Borel state spaces without terminal cost can be found in, e. g., (Hernández-Lerma and Lasserre, 1996).

2. Problem (7.6) can be modelled as an MDP with uncountable Borel state space $S \times \mathbb{R}$ and with finite disturbance space (cf. (Bertsekas and Shreve, 1978)). The system jumps from (s,ξ) to $(s,(\xi-c)/\beta)$ with probability $p^a_{ss'c}$ if action $a \in D(s)$ is chosen. At each time step, a cost of zero is incurred, and at the last time step, the terminal cost $v^*_{-1}(s,\xi) = \xi^-$, $(s,\xi) \in S \times \mathbb{R}$, has to be paid. Hence, theory of uncountable Borel state spaces, e. g., (Bertsekas and Shreve, 1978; Schäl, 2004), can be applied. By checking several assumptions, similar results as in Proposition 7.5.4 can be established with respect to the state space $S \times \mathbb{R}$.

Remark 7.5.6. 1. The optimal policy $\pi^{n,*}_\xi$ defined in the proof of Proposition 7.5.4 is deterministic. In general, $\pi^{n,*}_\xi$ is not Markovian since $\pi^{n,*}_\xi$ depends on the incurred costs.

2. The recursion formula of the deterministic optimal policy defined in (7.8) can be rewritten as

$$\left(\pi^{n+1,*}_\xi\right)_k(c_0,c_1,\ldots,c_{k-1},s') = \left(\pi^{n+1-k,*}_{\frac{\xi-c_0}{\beta^k}-\frac{c_1}{\beta^{k-1}}-\ldots-\frac{c_{k-1}}{\beta}}\right)_0(s') = \left(\pi^{n+1-k,*}_{\frac{1}{\beta^k}\left(\xi-\sum_{l=0}^{k-1}\beta^l c_l\right)}\right)_0(s'), \tag{7.9}$$
$$c_0, c_1, \ldots, c_{k-1} \in W,\ s' \in S,\ k = 1,\ldots,n+1.$$

So, the optimal action of the $(n+1)$-horizon problem with threshold ξ in k time steps after having paid the sequence of costs $c_0, c_1, \ldots, c_{k-1}$ at the preceding time steps 0, 1, …, $k-1$ is the same as in the $(n+1-k)$-horizon problem with threshold $\beta^{-k}\left(\xi-\sum_{l=0}^{k-1}\beta^l c_l\right)$ at the 0th time step. Moreover, it is remarkable that the optimal policy does not depend on the whole history of costs but only on the cumulated discounted cost $\sum_{l=0}^{k-1}\beta^l c_l$ together with the current system state and the current time step k as it can be seen from equation (7.9).

The next corollary links the optimal policy for problem (7.6) with policies which are optimal with respect to the expected total discounted cost criterion.

Corollary 7.5.7. *Assume the decision maker acts optimally for problem* (7.6).

1. *Let the incurred costs $c_0, \ldots, c_m \in W$ satisfy $\sum_{k=0}^m \beta^k c_k \geq \xi$ for some $m \in \{-1,0,\ldots,n-1\}$, then an optimal policy with respect to the expected total discounted cost is optimal for time steps l with $m < l \leq n$. (For $m=-1$, the assertion refers to the case $\xi \leq 0$.)*

2. *Let the incurred costs $c_0, \ldots, c_m \in W$ satisfy $\sum_{k=0}^m \beta^k c_k + \beta^{m+1}\frac{1-\beta^{n-m}}{1-\beta}U \leq \xi$ for some $m \in \{-1,0,\ldots,n-1\}$, then any action is optimal for time steps l with $m < l \leq n$. (For $m=-1$, the assertion refers to the case $\xi \geq (1-\beta^{n+1})U/(1-\beta)$.)*

Proof. To see 1, note that for $m \in \{0,\ldots,n-1\}$, $\pi \in \Pi$ and $s \in S$

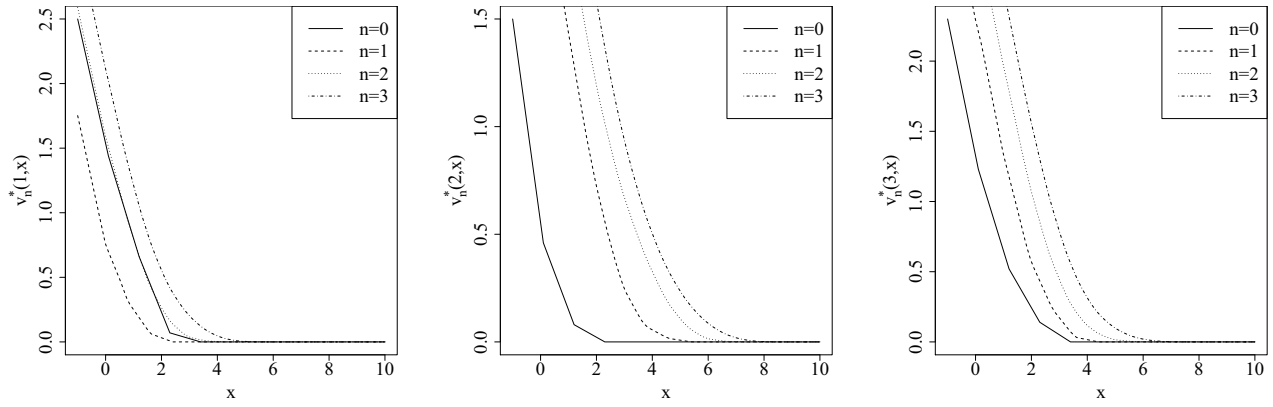
$$\begin{aligned} &E^\pi\left[(C^n-\xi)^+ \,\middle|\, C_k = c_k,\ k=0,\ldots,m,\ X_{m+1}=s\right] \\ &\quad = \sum_{k=0}^m \beta^k c_k - \xi + \beta^{m+1} E^\pi\left[C_{m+1}+\ldots+\beta^{n-(m+1)}C_n \,\middle|\, X_{m+1}=s\right], \end{aligned}$$

since $W \subset \mathbb{R}_{\geq 0}$ and therefore $\sum_{k=0}^m \beta^k c_k + \sum_{k=m+1}^n \beta^k C_k - \xi \geq 0$ P^π-a. s.. So, a standard expectation problem for the total discounted cost over an $(n-(m+1))$-horizon remains to be solved. For $m=-1$, we have $\xi \leq 0$ and $C^n - \xi \geq 0$ P^π-a. s. for every $\pi \in \Pi$ so that the positive part function is obsolete. Then the problem transfers into a standard expected total discounted cost problem. In the case of 2, we have for every $m \in \{-1,0,\ldots,n-1\}$ satisfying the assumption, for every $\pi \in \Pi$ and every $s \in S$

$$0 \leq E^\pi\left[(C^n-\xi)^+ \,\middle|\, C_k = c_k,\ k=0,\ldots,m,\ X_{m+1}=s\right] \leq \left(\sum_{k=0}^m \beta^k c_k + \beta^{m+1}\frac{1-\beta^{n-m}}{1-\beta}U - \xi\right)^+ = 0.$$

Hence, every action is optimal after time step m. □

So, the corollary tells us, how the decision maker has to act if the costs she has to pay until time step m are too high: in this case, she should act according to an optimal policy derived from the expected total discounted cost criterion for the



(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.1.: Value function of the example in section 7.5.3.

remaining time steps. Moreover, the decision maker can lean back if the incurred costs are so low until time step m that they satisfy 2 from Corollary 7.5.7 because any action choice after time step m is optimal.

The next statements give properties of the value functions $v^\pi(s,\cdot)$ and the optimal value function $v_n^*(s,\cdot)$ for fixed $s \in S$. We see that the value functions as well as the optimal value functions have a nice structure, which is exploited in the following section. In Figure 7.1, the value functions v_n^* for $n = 0,1,2,3$ of the example given in section 7.5.3 are illustrated. Note that there are non-negative terminal costs included in the example the cost distributions of which differ from those when state transitions occur. But in this case, subsequent statements are true, too (cf. Remark 7.5.13.2).

Lemma 7.5.8. *Let $n \in \mathbb{N}_0$, $\pi \in \Pi$ and $s \in S$. Then $v_n^\pi(s,\cdot)$ and $v_n^*(s,\cdot)$ are decreasing.*

Proof. Let $\xi \leq \eta$. Since $(\cdot)^+$ is increasing, it holds

$$v_n^\pi(s,\xi) = E^\pi\left[(C^n - \xi)^+ \,\middle|\, X_0 = s\right] \geq E^\pi\left[(C^n - \eta)^+ \,\middle|\, X_0 = s\right] = v_n^\pi(s,\eta).$$

Taking the infimum over all $\pi \in \Pi$ concludes the proof. □

Proposition 7.5.9. *Let $n \in \mathbb{N}_0$, $\pi \in \Pi$ and $s \in S$. Then $v_n^\pi(s,\cdot)$ is Lipschitz continuous with Lipschitz constant 1.*

Proof. This is an immediate consequence of the Lipschitz continuity of the positive part function $(\cdot)^+$: let $\xi \leq \eta$. Then we have for $n \in \mathbb{N}_0$, $\pi \in \Pi$ and $s \in S$

$$\begin{aligned}
|v_n^\pi(s,\xi) - v_n^\pi(s,\eta)| &\overset{\text{Lemma 7.5.8}}{=} v_n^\pi(s,\xi) - v_n^\pi(s,\eta) = E^\pi\left[(C^n - \xi)^+ - (C^n - \eta)^+ \,\middle|\, X_0 = s\right] \\
&= E^\pi\left[(C^n - \eta + \eta - \xi)^+ - (C^n - \eta)^+ \,\middle|\, X_0 = s\right] \overset{\text{Lemma A.3}}{\leq} E^\pi\left[(C^n - \eta)^+ + \eta - \xi - (C^n - \eta)^+ \,\middle|\, X_0 = s\right] \\
&= \eta - \xi = 1 \cdot |\xi - \eta|,
\end{aligned}$$

which is the assertion. □

A similar statement is true for the optimal value functions $v_n^*(s,\cdot)$ for fixed $s \in S$. Let us note this statement in the following corollary.

Corollary 7.5.10. *Let $n \in \mathbb{N}_0$ and $s \in S$. Then $v_n^*(s,\cdot)$ is Lipschitz continuous with Lipschitz constant 1.*

Proof. Let $\xi \leq \eta$. Then together with Lemma 7.5.8 we get

$$\begin{aligned}
|v_n^*(s,\xi) - v_n^*(s,\eta)| &= \left|\inf_{\pi\in\Pi} v_n^\pi(s,\xi) - \inf_{\pi\in\Pi} v_n^\pi(s,\eta)\right| \overset{\text{Lemma A.5}}{\leq} \sup_{\pi\in\Pi} |v_n^\pi(s,\xi) - v_n^\pi(s,\eta)| \\
&\overset{\text{Proposition 7.5.9}}{\leq} \sup_{\pi\in\Pi} |\xi - \eta| = 1 \cdot |\xi - \eta|,
\end{aligned}$$

since $v_n^\pi(s,\cdot) \geq 0$ for every $s \in S$ and $\pi \in \Pi$. □

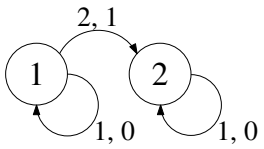


Figure 7.2.: MDP model of Example 7.5.12.

Another remarkable property of the value function $v_n^\pi(s,\cdot)$ is convexity. The same is not true for the optimal value function in general since the infimum of convex functions need not be convex as one can see from an example of two intersecting linear functions. This property is not needed in the further text. Nevertheless, it is proven in the next proposition.

Proposition 7.5.11. *Let $n \in \mathbb{N}_0$, $\pi \in \Pi$ and $s \in S$. Then $v_n^\pi(s,\cdot)$ is convex.*

Proof. This result follows in a similar way as Proposition 7.5.9 from the convexity of $(\cdot)^+$: let $n \in \mathbb{N}_0$, $\pi \in \Pi$, $s \in S$, ξ, $\eta \in \mathbb{R}$ and $\lambda \in (0,1)$. Then

$$\begin{aligned} v_n^\pi(s, \lambda\,\xi + (1-\lambda)\,\eta) &= E^\pi\left[\left(C^n - (\lambda\,\xi + (1-\lambda)\,\eta)\right)^+ \,\middle|\, X_0 = s\right] \\ &= E^\pi\left[\left(\lambda\,(C^n - \xi) + (1-\lambda)\,(C^n - \eta)\right)^+ \,\middle|\, X_0 = s\right] \\ &\leq \lambda\, E^\pi\left[(C^n - \xi)^+ \,\middle|\, X_0 = s\right] + (1-\lambda)\, E^\pi\left[(C^n - \eta)^+ \,\middle|\, X_0 = s\right] = \lambda\, v_n^\pi(s,\xi) + (1-\lambda)\, v_n^\pi(s,\eta), \end{aligned}$$

which is the definition of convexity. □

Up to this point, we established a backward induction algorithm for problem (7.6) for finite-horizon MDPs in Proposition 7.5.4. If the criterion is to minimize the expected total discounted cost over a finite horizon, a very similar backward induction algorithm also known as the dynamic programming algorithm holds (cf. (Bertsekas, 2005), Proposition 1.3.1). The difference between those two algorithms is that the backward step includes the outcomes of the states and additionally the incurred costs in the case of the average value-at-risk criterion, whereas for the expected costs only the outcomes of the states are relevant. Furthermore, the criterion considered in this section lacks the principle of optimality within the class of deterministic Markovian policies. Let $\pi^* = (\pi_0^*, \ldots, \pi_n^*, \ldots)$ be optimal for the n-horizon problem. Then for some $0 < m \leq n$, the policy $(\pi_m^*, \ldots, \pi_n^*, \ldots)$ need not be optimal for the $(n-m)$-horizon problem. This is demonstrated in the following example.

Example 7.5.12. Let $S = \{1,2\}$, $A = \{1,2\}$ with $D(1) = A$ and $D(2) = \{1\}$ and $W = \{0,1\}$ where the transition probabilities are given by

$$p_{110}^1 = 1, \quad p_{121}^2 = 1, \quad p_{220}^1 = 1,$$

and all remaining transition probabilities are zero. A sketch of this model is given in Figure 7.2 where the numbers on the arrows indicate the action and the corresponding deterministic costs. Let the discount factor be $\beta = 0.4$. In this example, we consider the 0- and 1-horizon case. In state 2, there is nothing to decide. Hence, we must only consider the following three deterministic Markovian policies π^1, π^2 and π^3 defined by

$$\begin{aligned} &\pi_0^1(1) = 1, \quad \pi_1^1(1) = 1, \\ &\pi_0^2(1) = 1, \quad \pi_1^2(1) = 2, \\ &\pi_0^3(1) = 2. \end{aligned}$$

Then we compute for $\xi = 0.5$

$$\begin{aligned} E^{\pi^1}\left[\left(C^1 - \xi\right)^+ \,\middle|\, X_0 = 1\right] &= (0 + 0 - \xi)^+ = 0, \\ E^{\pi^2}\left[\left(C^1 - \xi\right)^+ \,\middle|\, X_0 = 1\right] &= (0 + \beta - \xi)^+ = 0, \\ E^{\pi^3}\left[\left(C^1 - \xi\right)^+ \,\middle|\, X_0 = 1\right] &= (1 + 0 - \xi)^+ = 0.5. \end{aligned}$$

Hence, policies π^1 and π^2 both are optimal in the 1-horizon case. But in the 0-horizon case, we have for the truncated policies $\pi^{1,0} := (\pi_1^1, \ldots)$ and $\pi^{2,0} := (\pi_1^2, \ldots)$

$$E^{\pi^{1,0}}\left[\left(C^0 - \xi\right)^+ \,\middle|\, X_0 = 1\right] = (0 - \xi)^+ = 0,$$

$$E^{\pi^{2,0}}\left[\left(C^0-\xi\right)^+\,\middle|\,X_0=1\right]=(1-\xi)^+=0.5.$$

Therefore, the principle of optimality does not hold for policy π^2 since it is not optimal for the 0-horizon problem although it is optimal in the 1-horizon case. This is because after having paid the current cost, in general, the threshold value ξ has to be adjusted appropriately to obtain an optimal policy according to equation (7.9).

We conclude this section with a remark.

Remark 7.5.13. 1. In this chapter, we assumed $W\subset\mathbb{R}_{\geq 0}$. But Lemmas 7.5.1–7.5.8, Propositions 7.5.2–7.5.9 and Corollary 7.5.10 hold true for any finite $W\subset\mathbb{R}$. Corollary 7.5.7 may fail in the general case since the costs C^n may not be increasing under all policies. But this lack can be fixed by assigning appropriate bounds including $\min W$.

2. We could also include positive terminal costs in the model. That is, in the last step the decision maker has to pay a cost according to a distribution which depends on the current state but differs from the cost distribution when there are state transitions, here according to $p^a_{sc}=P(C_n=c\,|\,X_n=s,A_n=a)$, $s\in S$, $c\in W$, $a\in D(s)$. Then the backward induction algorithm of Proposition 7.5.4 has the form

$$\begin{aligned}
&v^*_0=T_0v^*_{-1},\\
&v^*_n=T^{n-1}T_0v^*_{-1},\quad\text{where}\\
&T_0v(\xi)\coloneqq\beta\min_{a\in D(s)}\sum_{c_0\in W}p^a_{sc_0}\,v\left(\frac{\xi-c_0}{\beta}\right),\quad s\in S,\,\xi\in\mathbb{R},\,v:\mathbb{R}\to\mathbb{R},\quad\text{and}\\
&v^*_{-1}(s,\xi)\coloneqq\xi^-,\quad\xi\in\mathbb{R}.
\end{aligned}$$

The preceding proofs can easily be adapted to this formulation.

7.5.2. The Average Value-at-Risk Criterion

Having solved the intermediate problem (7.6) for every $\xi\in\mathbb{R}$, we can now return to the original problem given in (7.2), which is to find a policy that minimizes the average value-at-risk for a finite horizon. We can restate it in the following form. For some given confidence level $\tau\in(0,1)$ and initial state $x_0\in S$, we wish to solve:

$$\text{Minimize }\xi+\frac{1}{1-\tau}\,v^*_n(x_0,\xi)\text{ over all }\xi\in\mathbb{R}. \tag{7.10}$$

The functions $v^*_n(x_0,\cdot)$ are assumed to be known for all $x_0\in S$, although their computation may be hard. Define w^τ_n as the objective function of problem (7.10), that is

$$w^\tau_n(x_0,\xi)\coloneqq\xi+\frac{1}{1-\tau}\,v^*_n(x_0,\xi),\quad x_0\in S,\,\xi\in\mathbb{R}.$$

In the following, we study the behaviour of w^τ_n. From this, we can finally conclude that there exists a deterministic optimal policy for problem (7.2) (cf. Theorem 7.5.17).

Lemma 7.5.14. *The function w^τ_n is Lipschitz continuous with Lipschitz constant $1+1/(1-\tau)=(2-\tau)/(1-\tau)$.*

Proof. This is a consequence of Corollary 7.5.10 and Lemma A.6 since the identity mapping on $\mathbb{R}$ is Lipschitz continuous with Lipschitz constant 1. □

Lemma 7.5.15. *For every $s\in S$, it holds*

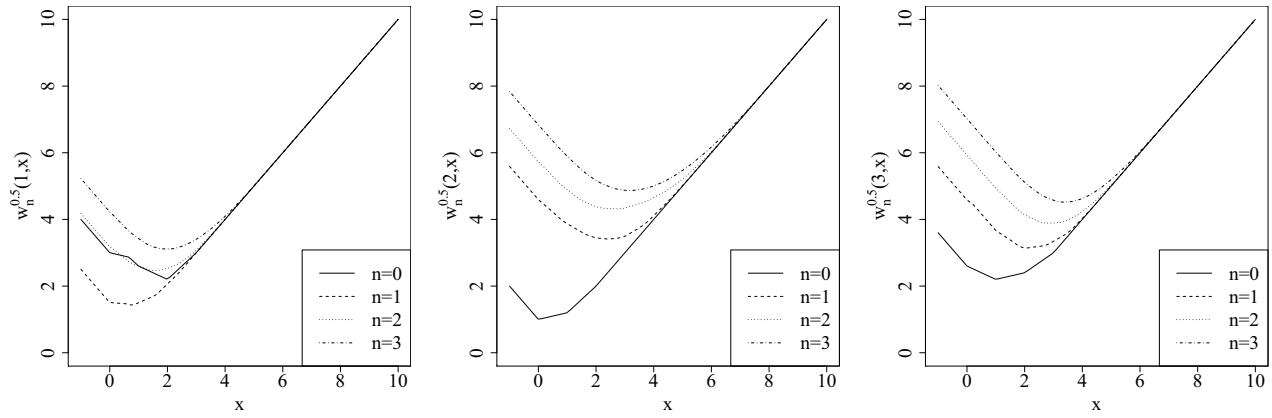
1. $w^\tau_n(s,\xi)\overset{\xi\to\infty}{\to}\infty$ *and*
2. $w^\tau_n(s,\xi)\overset{\xi\to-\infty}{\to}\infty$.

Proof. To obtain 1, consider $\xi\geq U/(1-\beta)$. Then for $s\in S$, since $C^n\leq\xi$ P^π-a. s. for every $\pi\in\Pi$,

$$w^\tau_n(s,\xi)=\xi+\frac{1}{1-\tau}\inf_{\pi\in\Pi}E^\pi\left[\left(C^n-\xi\right)^+\,\middle|\,X_0=s\right]=\xi\to\infty,\quad\xi\to\infty.$$

To see 2, let $\xi\leq 0$. Since $C^n\geq 0$ P^π-a. s. under every policy $\pi\in\Pi$, we obtain

$$v^*_n(s,\xi)=\inf_{\pi\in\Pi}E^\pi\left[\left(C^n-\xi\right)^+\,\middle|\,X_0=s\right]=\inf_{\pi\in\Pi}E^\pi\left[C^n\,|\,X_0=s\right]-\xi,\quad s\in S.$$



(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.3.: Function $w_n^{0.5}$ of the example in section 7.5.3.

Hence, for $s \in S$,

$$w_n^\tau(s,\xi) = \frac{1}{1-\tau} \inf_{\pi\in\Pi} E^\pi\left[C^n \mid X_0 = s\right] + \left(1 - \frac{1}{1-\tau}\right)\xi = \frac{1}{1-\tau} \inf_{\pi\in\Pi} E^\pi\left[C^n \mid X_0 = s\right] - \frac{\tau}{1-\tau}\xi \to \infty, \quad \xi \to -\infty,$$

since $\tau/(1-\tau) > 0$. □

In Figure 7.3, the functions $w_n^{0.5}$ are illustrated for $n = 0,1,2,3$ for the numerical example of section 7.5.3, from which we can see the continuity and the increase towards ∞ for $\xi \to \pm\infty$ for the functions $w_n^{0.5}(x_0,\xi)$ for all $x_0 \in S$ proven in Lemmas 7.5.14 and 7.5.15. The next proposition makes the almost clear statement that such a function attains its minimum.

Proposition 7.5.16. *There exists a solution to problem* (7.10)*.*

Proof. We have to show that there is an $x^* \in \mathbb{R}$ such that $w_n^\tau(x_0,\xi^*) = \inf_{\xi\in\mathbb{R}} w_n^\tau(x_0,\xi)$. Because $w_n^\tau(x_0,\xi)$ is growing above all bounds for $\xi \to \pm\infty$ (cf. Lemma 7.5.15), there exists some $R \in \mathbb{R}$ such that $K := \{\xi \in \mathbb{R} : w_n^\tau(x_0,\xi) \le R\} \neq \emptyset$ and bounded. Furthermore, K is closed since $w_n^\tau(x_0,\cdot)$ is continuous by Lemma 7.5.14. Hence, K is compact. So, by a theorem of Weierstraß (cf. Theorem 5.2.12 in (Trench, 2009)), there is some $\xi^* \in K$ such that $w_n^\tau(x_0,\xi^*) = \min_{\xi\in K} w_n^\tau(x_0,\xi)$. Since $w_n^\tau(x_0,\xi) > w_n^\tau(x_0,\xi^*)$ for all $\xi \notin K$, the assertion follows. □

With Proposition 7.5.16 in hand, we can prove the existence of a deterministic optimal policy for the AV@R$_\tau$-problem (7.2).

Theorem 7.5.17. *There exists a deterministic optimal policy for problem* (7.2)*.*

Proof. Let $\xi^* \in \mathbb{R}$ be a solution to problem (7.10). Define a deterministic optimal policy $\pi_{\xi^*}^{n,*} \in \Pi^d$ for the intermediate problem (7.6) with $\xi = \xi^*$ and horizon $n \in \mathbb{N}_0$ according to (7.8). Then $\pi_{\xi^*}^{n,*}$ is AV@R$_\tau$-optimal since

$$\begin{aligned}
\inf_{\pi\in\Pi} \mathrm{AV@R}_\tau^\pi\left(C^n \mid X_0 = x_0\right) &\overset{\text{Theorem 6.2.3}}{=} \inf_{\pi\in\Pi}\left\{\inf_{\xi\in\mathbb{R}}\left\{\xi + \frac{1}{1-\tau}E^\pi\left[\left(C^n-\xi\right)^+ \,\middle|\, X_0 = x_0\right]\right\}\right\} \\
&\overset{\text{Lemma A.4}}{=} \inf_{\xi\in\mathbb{R}}\left\{\xi + \frac{1}{1-\tau}\underbrace{\inf_{\pi\in\Pi} E^\pi\left[\left(C^n-\xi\right)^+ \,\middle|\, X_0 = x_0\right]}_{=v_n^*(x_0,\xi)}\right\} \\
&\overset{\text{Proposition 7.5.16}}{=} \xi^* + \frac{1}{1-\tau}\inf_{\pi\in\Pi} E^\pi\left[\left(C^n-\xi^*\right)^+ \,\middle|\, X_0 = x_0\right] \overset{\text{Proposition 7.5.4}}{=} \xi^* + \frac{1}{1-\tau}E^{\pi_{\xi^*}^{n,*}}\left[\left(C^n-\xi^*\right)^+ \,\middle|\, X_0 = x_0\right] \\
&\ge \inf_{\xi\in\mathbb{R}}\left\{\xi + \frac{1}{1-\tau}E^{\pi_{\xi^*}^{n,*}}\left[\left(C^n-\xi\right)^+ \,\middle|\, X_0 = x_0\right]\right\} \overset{\text{Theorem 6.2.3}}{=} \mathrm{AV@R}_\tau^{\pi_{\xi^*}^{n,*}}\left(C^n \mid X_0 = x_0\right).
\end{aligned}$$

Hence $\mathrm{AV@R}_\tau^{\pi_{\xi^*}^{n,*}}\left(C^n \mid X_0 = x_0\right) = \inf_{\pi\in\Pi} \mathrm{AV@R}_\tau^\pi\left(C^n \mid X_0 = x_0\right)$. □

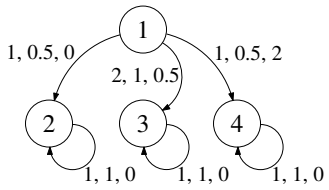


Figure 7.4.: MDP model of Example 7.5.19.

We conclude this section with some remarks and two small examples.

Remark 7.5.18.

1. The value ξ^* in the proof of Theorem 7.5.17 is almost the value-at-risk of C^n under policy $\pi^{n,*}_{\xi^*}$. If ξ^* is the smallest value such that it solves problem (7.10), then indeed ξ^* is the value-at-risk of C^n under policy $\pi^{n,*}_{\xi^*}$ by Theorem 6.2.3. But it need not be the case that $\pi^{n,*}_{\xi^*}$ minimizes the value-at-risk of C^n, i. e., a policy which is optimal with respect to the average value-at-risk criterion for level $\tau \in (0,1)$ need not be optimal with respect to the value-at-risk criterion for level τ. This is illustrated in Example 7.5.19.

2. Since W is finite, the total discounted cost C^n can only take values from the set $W^n := \left\{\sum_{k=0}^n \beta^k c_k : c_k \in W\right\}$, which is a discrete set. Therefore, for every policy only values lying in W^n come into consideration as the value-at-risk of C^n. Since the value-at-risk of an AV@R$_\tau$-optimal policy minimizes problem (7.10), one has to consider $v_n^*(x_0,\xi)$ for $\xi \in W^n$ only, which reduces the complexity of the problem considerably.

3. Theorem 7.5.17 is also true for arbitrary finite $W \subset \mathbb{R}$. Only the proof of Lemma 7.5.15.2 has to be changed: one has to consider $\xi \leq L/(1-\beta)$ where $L := \min W$. Since the costs C^n cannot drop below $L/(1-\beta)$ and hence $C^n - \xi \geq 0$. Together with Remark 7.5.13.1, the results of this section follow.

4. In a similar manner, we can include positive terminal costs, leading to the same results of this section.

5. The deterministic optimal policy of Theorem 7.5.17 has the following form: at first, the decision maker has to determine the appropriate threshold value $\xi^* \in \mathbb{R}$ which she should not exceed in order to minimize the average value-at-risk by solving problem (7.10). After that, she acts optimally with respect to the inner problem (7.6) with threshold value ξ^*. After each step, the decision maker has to adjust her preference, which $\xi \in \mathbb{R}$ she should not exceed in the remaining steps according to equation (7.9).

6. The problem of minimizing AV@R$^\pi_\tau(C^n \,|\, X_0 = x_0)$ we posed does not satisfy the principle of optimality within the class of deterministic Markovian policies. Having found some $\xi^*(x_0,\tau)$ which should not be exceeded in order to minimize the average value-at-risk at level τ, problem (7.6) with $\xi = \xi^*(x_0,\tau)$ might be time-inconsistent as we have seen in Example 7.5.12. To make things explicit, Example 7.5.20 demonstrates this property by direct computation.

Example 7.5.19. Here, we shortly illustrate that a general AV@R$_\tau$-optimal policy might not be V@R$_\tau$-optimal. The MDP model is the following: $S = \{1,2,3,4\}$, $A = \{1,2\}$, $D(1) = A$, $D(2) = D(3) = D(4) = \{1\}$ and $W = \{0, 0.5, 2\}$ with transition probabilities

$$p^1_{120} = 0.5, \quad p^1_{142} = 0.5, \quad p^2_{13(0.5)} = 1.$$

A sketch of this model can be found in Figure 7.4 where the numbers on the arrows denote the action, the transition probability and the cost respectively. Let $\tau = 0.5$, β be arbitrary, and let the initial state be $x_0 = 1$. Consider the 0-horizon problem, i. e., the decision maker has to make exactly one decision. As we have shown, there is a deterministic optimal policy to the AV@R$_\tau$-criterion. Consider the two only deterministic policies, which are defined by $\pi_1(1) := 1$ and $\pi_2(1) := 2$. Then we have

$$\text{AV@R}^{\pi_1}_{0.5}(C^0 \,|\, X_0 = 1) = 2 \quad \text{and} \quad \text{AV@R}^{\pi_2}_{0.5}(C^0 \,|\, X_0 = 1) = 0.5.$$

So, π_2 is AV@R$_{0.5}$-optimal. But for the value-at-risk at level 0.5 of C^0 under π_1 and under π_2 respectively we have

$$\text{V@R}^{\pi_1}_{0.5}(C^0 \,|\, X_0 = 1) = 0 \quad \text{and} \quad \text{V@R}^{\pi_2}_{0.5}(C^0 \,|\, X_0 = 1) = 0.5.$$

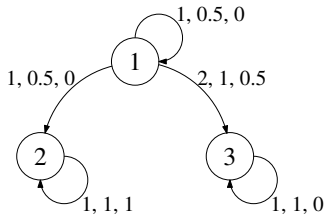


Figure 7.5.: MDP model of Example 7.5.20.

Example 7.5.20. In this example, we demonstrate that the principle of optimality does not hold for the average value-at-risk criterion for deterministic Markovian policies. Let $S = \{1,2,3\}$, $A = \{1,2\}$, $D(1) = A$, $D(2) = D(3) = \{1\}$ and $W = \{0, 0.5, 1\}$ with transition probabilities

$$p^1_{110} = 0.5, \quad p^1_{120} = 0.5, \quad p^2_{13(0.5)} = 1, \quad p^1_{221} = 1, \quad p^1_{330} = 1.$$

A sketch of this model can be found in Figure 7.5. Further, let $\tau = 0.5$ and $\beta = 0.4$. Let us again consider the 0- and the 1-horizon problem for the initial state 1. By Proposition 7.5.17, we have to consider deterministic policies only. There are three deterministic policies since there is nothing to decide in states 2 and 3. Define the policies $\pi^1 = (\pi^1_0, \pi^1_1, \ldots)$, $\pi^2 = (\pi^2_0, \pi^2_1, \ldots)$ and $\pi^3 = (\pi^3_0, \pi^3_1, \ldots)$ by

$$\begin{aligned}
&\pi^1_0(1) = 1, \quad \pi^1_1(1) = 1,\\
&\pi^2_0(1) = 1, \quad \pi^2_1(1) = 2,\\
&\pi^3_0(1) = 2.
\end{aligned}$$

Then we compute the following conditional distributions

$$\begin{aligned}
&P^{\pi^1}\left(C^1 = 0 \,\middle|\, X_0 = 1\right) = 0.5, \qquad P^{\pi^1}\left(C^1 = 0.4 \,\middle|\, X_0 = 1\right) = 0.5,\\
&P^{\pi^2}\left(C^1 = 0.2 \,\middle|\, X_0 = 1\right) = 0.5, \qquad P^{\pi^2}\left(C^1 = 0.4 \,\middle|\, X_0 = 1\right) = 0.5,\\
&P^{\pi^3}\left(C^1 = 0.5 \,\middle|\, X_0 = 1\right) = 1,
\end{aligned}$$

from which we obtain

$$\begin{aligned}
&\mathrm{AV@R}^{\pi^1}_{0.5}\left(C^1 \,\middle|\, X_0 = 1\right) = 0.4,\\
&\mathrm{AV@R}^{\pi^2}_{0.5}\left(C^1 \,\middle|\, X_0 = 1\right) = 0.4,\\
&\mathrm{AV@R}^{\pi^3}_{0.5}\left(C^1 \,\middle|\, X_0 = 1\right) = 0.5.
\end{aligned}$$

Hence, the two policies π^1 and π^2 are optimal within the class of deterministic Markovian policies in the 1-horizon case. But for the truncated policies $\pi^{1,0} = (\pi^1_1, \ldots)$ and $\pi^{2,0} = (\pi^2_1, \ldots)$, we have

$$\begin{aligned}
&\mathrm{AV@R}^{\pi^{1,0}}_{0.5}\left(C^0 \,\middle|\, X_0 = 1\right) = 0,\\
&\mathrm{AV@R}^{\pi^{2,0}}_{0.5}\left(C^0 \,\middle|\, X_0 = 1\right) = 0.5
\end{aligned}$$

and $\pi^{2,0}$ is not optimal in the 0-horizon case, which shows that the principle of optimality does not hold for the average value-at-risk criterion within the class of deterministic Markovian policies.

In (Boda and Filar, 2006), this time inconsistency of the average value-at-risk with respect to decision rules is demonstrated by another example. It concerns portfolio optimization and is more complex than the preceding example since the values of the underlying assets follow continuous distributions.

7.5.3. Numerical Example

In this section, we examine a small example. We suppose that there are three states, in which one can choose between three actions. At each time step, the decision maker has to pay a cost lying in $W := \{0,1,2,3\}$. A list of all parameters of the model can be found in section B.3. We consider a horizon of length 1, i. e., the decision maker has to make two

τ	x_0	π	$\widehat{\text{AV@R}}_\tau^\pi$	τ	x_0	π	$\widehat{\text{AV@R}}_\tau^\pi$	τ	x_0	π	$\widehat{\text{AV@R}}_\tau^\pi$
0.1	1	opt_0.1	0.8396	0.5	1	opt_0.1	1.4596	0.95	1	opt_0.1	2.4000
0.1	1	opt_0.5	0.8685	0.5	1	opt_0.5	1.4243	0.95	1	opt_0.5	2.4000
0.1	1	opt_0.95	0.9460	0.5	1	opt_0.95	1.5624	0.95	1	opt_0.95	2.4000
0.1	1	opt_0.95_2	1.7518	0.5	1	opt_0.95_2	2.1939	0.95	1	opt_0.95_2	2.4000
0.1	1	random	2.2801	0.5	1	random	3.1455	0.95	1	random	4.9795
0.1	2	opt_0.1	2.5492	0.5	2	opt_0.1	3.4363	0.95	2	opt_0.1	4.9123
0.1	2	opt_0.5	2.5927	0.5	2	opt_0.5	3.4221	0.95	2	opt_0.5	4.9392
0.1	2	opt_0.95	2.6782	0.5	2	opt_0.95	3.4403	0.95	2	opt_0.95	4.9110
0.1	2	opt_0.95_2	2.8083	0.5	2	opt_0.95_2	3.6682	0.95	2	opt_0.95_2	4.9138
0.1	2	random	3.4530	0.5	2	random	4.3717	0.95	2	random	5.4000
0.1	3	opt_0.1	2.4759	0.5	3	opt_0.1	3.1443	0.95	3	opt_0.1	4.4000
0.1	3	opt_0.5	2.4955	0.5	3	opt_0.5	3.1312	0.95	3	opt_0.5	4.4000
0.1	3	opt_0.95	2.5847	0.5	3	opt_0.95	3.2869	0.95	3	opt_0.95	4.2184
0.1	3	opt_0.95_2	2.8918	0.5	3	opt_0.95_2	3.3905	0.95	3	opt_0.95_2	4.2408
0.1	3	random	3.0854	0.5	3	random	3.8869	0.95	3	random	5.4000

Table 7.1.: Results after 100,000 simulation runs.

decisions, one at time step 0 and another one at time step 1. Further, we assume that there are terminal costs according to distributions for the different actions which differ from the transition probabilities at time step 0. Note that the results for the foregoing model apply also in this case by Remark 7.5.18.4.

Optimal policies for the AV@R$_\tau$-criterion were computed for the confidence levels $\tau = 0.1, 0.5, 0.95$. For $\tau = 0.95$, we defined two distinct optimal policies. The optimal policies were derived according to Theorem 7.5.17: we determined extreme points of $w_1^\tau(x_0,\cdot)$ and then defined optimal deterministic policies as given in (7.8) with the aim not to exceed the respective extreme points. These policies only differ at the second decision making step. At time step 0, all policies choose the same actions. To compare the results, we also implemented a random policy, which always chooses with uniform probability one of the three actions.

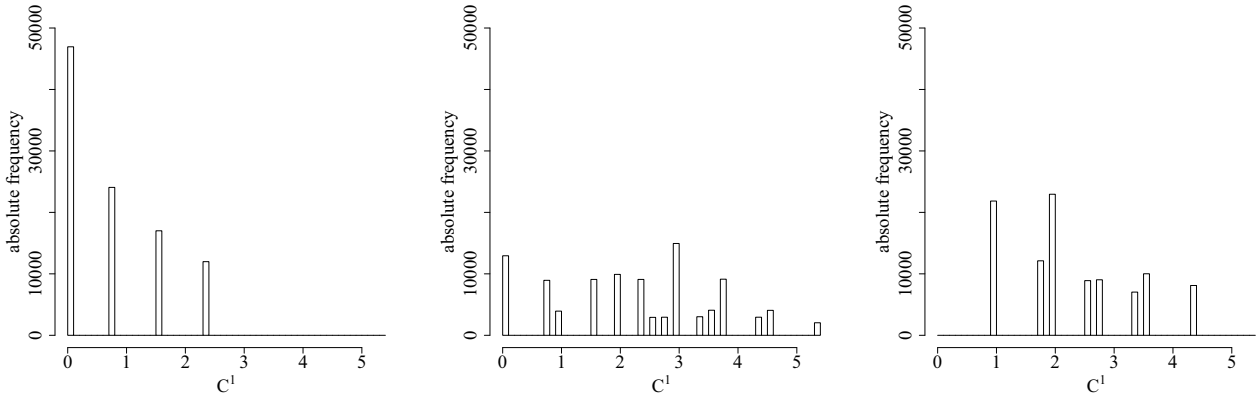
It is an interesting point that for policy `opt_0.1` for all time steps and for all states the actions are uniquely determined, whereas there are some histories for which the action can be chosen arbitrarily for an AV@R$_{0.5}$-optimal policy and in even more states for AV@R$_{0.95}$-optimal policies. The reason is rather simple: in this example, we have that with increasing τ the threshold value $\xi^* = \xi^*(x_0,\tau)$ for the intermediate criterion (cf. Theorem 7.5.17) increases, too, as one can see from Figures 7.1 and 7.3. Hence $(\xi^*(x_0,\tau) - c_0)/\beta$ increases. From this, we obtain that the set $\{c_0 \in W : (\xi^*(x_0,\tau) - c_0)/\beta \geq 3\}$ increases if τ increases. If $(\xi^*(x_0,\tau) - c_0)/\beta \geq 3$, then every action is optimal. This behaviour can be seen from policy `opt_0.95` and `opt_0.95_2` for state 1. Both policies are optimal for level 0.95 but policy `opt_0.95_2` chooses actions which lead to a higher cost at times step 1 such that the total discounted cost still remains below the value-at-risk at level 0.95 of this policy, and hence, it does not have an impact on the average value-at-risk.

Since exact computation of the average value-at-risk is quite exhausting, the five policies were simulated 100,000 times for each initial state to estimate the respective average-value-at-risk. The total discounted cost of the ith run is C_i^1, $i = 1, \ldots, 100{,}000$, for a fixed $x_0 \in S$. As an estimator for the average value-at-risk under policy π, we used

$$\widehat{\text{AV@R}}_\tau^\pi(C^1 \,|\, X_0 = x_0) = \frac{\sum_{i=100{,}000\,\tau+1}^{100{,}000} C_{(i)}^1}{100{,}000\,(1-\tau)},$$

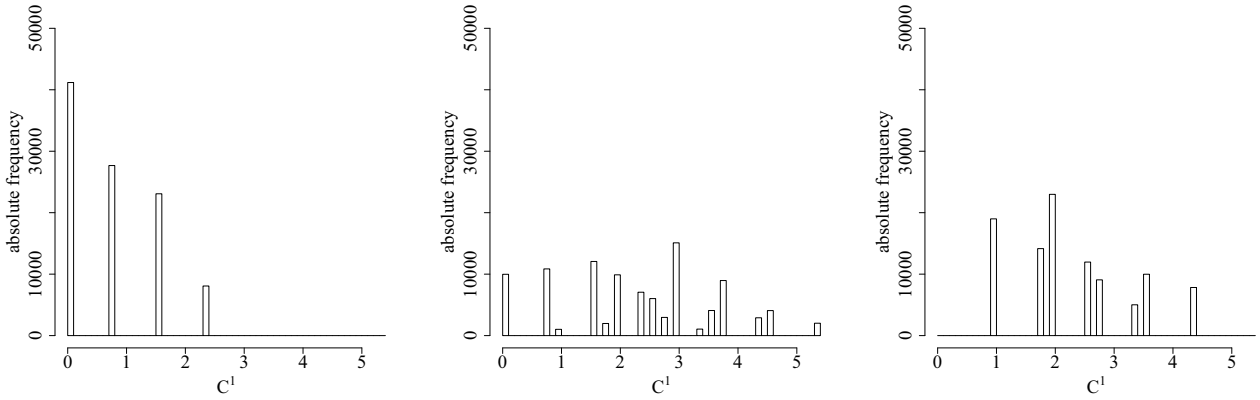
where $C_{(i)}^1$ is the ith largest of the C_j^1, $j = 1, \ldots, 100{,}000$, i. e., $(C_{(i)}^1)_{i=1,\ldots,100{,}000}$ is the order statistics of $(C_i^1)_{i=1,\ldots,100{,}000}$. The results are demonstrated in Table 7.1 and Figures 7.6–7.10. From Table 7.1, it can be seen that the policies are optimal for the levels they supposed to be optimal for, except for small deviations for level 0.95.

Note that for $x_0 = 1$ and $\tau = 0.95$, all policies except `random` seem to be AV@R$_{0.95}$-optimal. But constructing AV@R$_{0.95}$-optimal policies according to Theorem 7.5.17, where the respective threshold value is $\xi^*(1, 0.95) = 2.4$, leads to the optimal policies `opt_0.95` and `opt_0.95_2`. However, the policies `opt_0.1` and `opt_0.5` are no solutions to the intermediate problem for $\xi = \xi^*(1, 0.95)$. Note that this is not a contradiction to the results of the preceding sections. We proved that the construction of policies according to (7.8) leads to optimal policies. But we did not state that an optimal policy has to be based on such a threshold value $\xi^*(1, 0.95)$. There may be, as in this numerical example, optimal policies of another type.



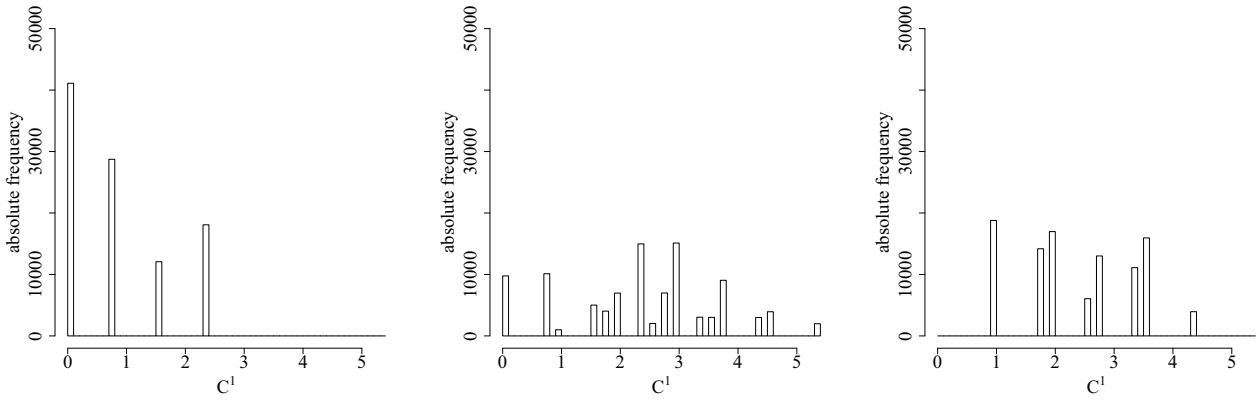
(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.6.: Histograms of the AV@$R_{0.1}$-optimal policy `opt_0.1`.



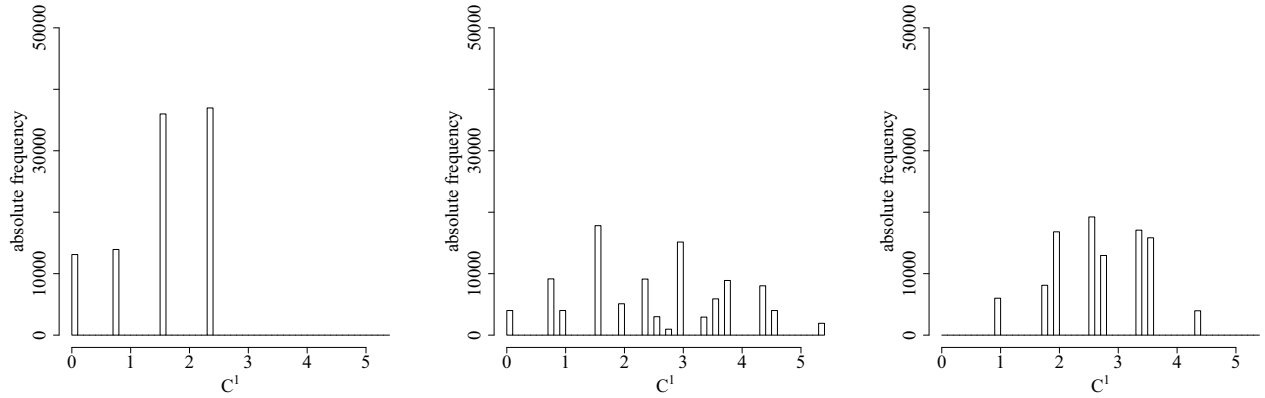
(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.7.: Histograms of the AV@$R_{0.5}$-optimal policy `opt_0.5`.



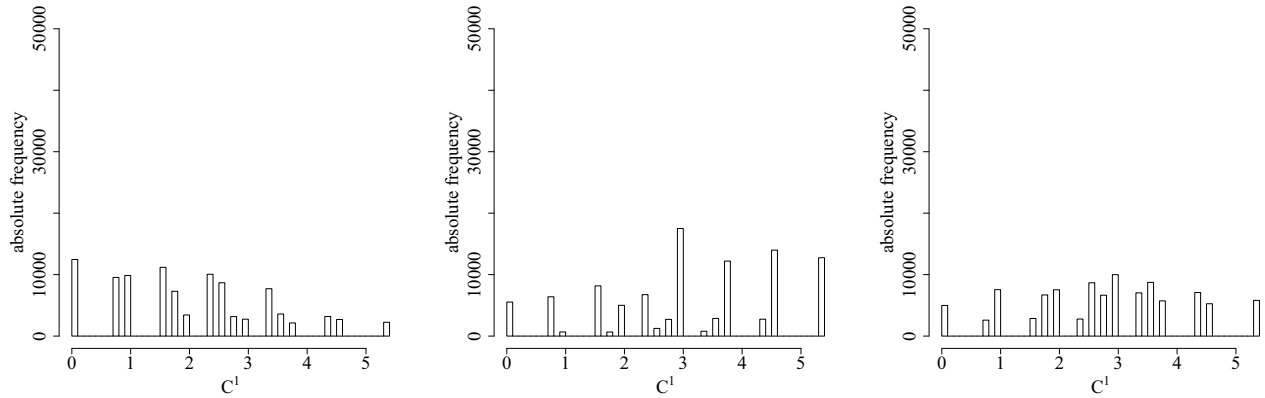
(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.8.: Histograms of the AV@$R_{0.95}$-optimal policy `opt_0.95`.



(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.9.: Histograms of the AV@$R_{0.95}$-optimal policy `opt_0.95_2`.



(a) Start in $x_0 = 1$. (b) Start in $x_0 = 2$. (c) Start in $x_0 = 3$.

Figure 7.10.: Histograms of the random policy `random`.

As a further example, we have a look at $x_0 = 3$ and $\tau = 0.95$. In this case, the policies `opt_0.95` and `opt_0.95_2` are rather bad for the AV@$R_{0.1}$- and for the AV@$R_{0.5}$-criterion. But they outperform the remaining policies for the AV@$R_{0.95}$-criterion. From the histograms, we see that there is a smaller probability for the policies `opt_0.95` and `opt_0.95_2` to end up with a cost of 4.4 than the other policies. Therefore, the cost 3.6 is included into the computation of the average value-at-risk. In contrast, the other policies tend to put high probability to lower costs together with the acceptance to increase the probability of higher costs.

7.6. The Infinite Horizon

In this section, we want to solve the average value-at-risk criterion for the total discounted cost if the time horizon is infinite, i. e., problem (7.3), which can be rewritten as in (7.5).

7.6.1. An Intermediate Criterion

At first, we consider the inner problem of (7.5) as we did for solving the finite-horizon case. We define the *infinite-horizon value function* for an arbitrary policy $\pi \in \Pi$ by

$$v^\pi(s,\xi) := E^\pi\left[(C^\infty - \xi)^+ \,\middle|\, X_0 = s\right], \quad s \in S,\ \xi \in \mathbb{R}.$$

For the given initial state $x_0 \in S$ and the threshold value $\xi \in \mathbb{R}$ the inner problem is:

$$\text{Minimize } v^\pi(x_0,\xi) \text{ over all } \pi \in \Pi. \tag{7.11}$$

Again, the inner problem is independent of the level τ. This is a problem for MDPs where the decision maker wants the total discounted cost not to exceed the threshold ξ. The total discounted costs the decision maker has to pay are penalized by the amount of excess. In the following, let

$$v^*(s,\xi) := \inf_{\pi\in\Pi} v^\pi(s,\xi), \quad s \in S,\ \xi \in \mathbb{R},$$

be the *optimal value function* for the infinite-horizon problem. To tackle the problem, we consider two distinct methods. First, we study the convergence of v_n^π, $\pi \in \Pi$, and v_n^* with results from the finite-horizon analysis, whereas we establish results by making use of the Banach fixed point theorem in the second approach.

Convergence Analysis

First, we study the convergence of v_n^π and v_n^* directly. This procedure leads to ε-optimal policies for problem (7.11) in a rather straightforward way.

Lemma 7.6.1. *Let $n \in \mathbb{N}_0$ and $\pi \in \Pi$. Then we have*

$$v_n^\pi \le v_{n+1}^\pi \quad \text{and} \quad v_n^* \le v_{n+1}^*.$$

Proof. Let $s \in S$ and $\xi \in \mathbb{R}$. Since $C_{n+1} \ge 0$ P^π-a. s. for every $\pi \in \Pi$, we get

$$\begin{aligned} v_{n+1}^\pi(s,\xi) &= E^\pi\left[\left(C_0 + \beta C_1 + \ldots + \beta^{n+1} C_{n+1} - \xi\right)^+ \,\middle|\, X_0 = s\right] \ge E^\pi\left[(C_0 + \beta C_1 + \ldots + \beta^n C_n - \xi)^+ \,\middle|\, X_0 = s\right] \\ &= v_n^\pi(s,\xi), \end{aligned}$$

and taking the infimum proves the second assertion. □

Since $0 \le v_n^\pi \le U/(1-\beta)$ for every $n \in \mathbb{N}_0$ and v_n^π is increasing in n by Lemma 7.6.1, we obtain that the pointwise limit $\lim_{n\to\infty} v_n^\pi =: v_\infty^\pi$ exists for every $\pi \in \Pi$. Furthermore, from $0 \le v_n^\pi \le U/(1-\beta)$ for all $n \in \mathbb{N}_0$, we get $0 \le v_\infty^\pi \le U/(1-\beta)$. The next proposition identifies v_∞^π with v^π.

Proposition 7.6.2. *For every $\pi \in \Pi$, it holds*

$$v_\infty^\pi = v^\pi.$$

Proof. Let $s \in S$ and $\xi \in \mathbb{R}$. Since $(C^\infty - \xi)^+ \geq 0$ and $(C^n - \xi)^+$ are increasing in n P^π-a. s. under $\pi \in \Pi$, an application of the monotone convergence theorem (cf. (Billingsley, 1995), Theorem 16.2) yields

$$v_n^\pi(s,\xi) = E^\pi\left[(C^n-\xi)^+ \,\middle|\, X_0 = s\right] \uparrow E^\pi\left[(C^\infty-\xi)^+ \,\middle|\, X_0 = s\right] = v^\pi(s,\xi), \quad n \to \infty.$$

On the other hand $v_n^\pi(s,\xi) \overset{n\to\infty}{\to} v_\infty^\pi(s,\xi)$ by definition. Thus, the assertion follows. □

The following lemma gives upper bounds for v^π and v^* in terms of the finite-horizon value functions v_n^π and v_n^* respectively, where the threshold ξ has to be modified for the finite-horizon problem in an appropriate manner.

Lemma 7.6.3. *Let $n \in \mathbb{N}_0$, $s \in S$, $\xi \in \mathbb{R}$ and $\pi \in \Pi$. Then*

$$v^\pi(s,\xi) \leq v_n^\pi\left(s, \xi - \frac{U\beta^{n+1}}{1-\beta}\right) \quad \text{and} \quad v^*(s,\xi) \leq v_n^*\left(s, \xi - \frac{U\beta^{n+1}}{1-\beta}\right).$$

Moreover, it holds

$$v^\pi(s,\xi) \leq v_n^\pi(s,\xi) + \frac{U\beta^{n+1}}{1-\beta} \quad \text{and} \quad v^*(s,\xi) \leq v_n^*(s,\xi) + \frac{U\beta^{n+1}}{1-\beta}.$$

Proof. The first assertion follows from the monotonicity of $(\cdot)^+$ since

$$\begin{aligned} v^\pi(s,\xi) &= E^\pi\left[(C^\infty-\xi)^+ \,\middle|\, X_0 = s\right] = E^\pi\left[\left(C^n + \beta^{n+1}\sum_{k=0}^\infty \beta^k C_{n+k+1} - \xi\right)^+ \,\middle|\, X_0 = s\right] \\ &\leq E^\pi\left[\left(C^n + \beta^{n+1}\frac{U}{1-\beta} - \xi\right)^+ \,\middle|\, X_0 = s\right] = E^\pi\left[\left(C^n - \left(\xi - \frac{U\beta^{n+1}}{1-\beta}\right)\right)^+ \,\middle|\, X_0 = s\right] \\ &= v_n^\pi\left(s, \xi - \frac{U\beta^{n+1}}{1-\beta}\right). \end{aligned}$$

The second assertion follows by taking the infimum over $\pi \in \Pi$. Similarly, we obtain

$$\begin{aligned} v^\pi(s,\xi) &\leq E^\pi\left[\left(C^n + \beta^{n+1}\frac{U}{1-\beta} - \xi\right)^+ \,\middle|\, X_0 = s\right] \overset{\text{Lemma A.3}}{\leq} E^\pi\left[(C^n-\xi)^+ \,\middle|\, X_0 = s\right] + \frac{U\beta^{n+1}}{1-\beta} \\ &= v_n^\pi(s,\xi) + \frac{U\beta^{n+1}}{1-\beta}. \end{aligned}$$

The second part follows by taking the infimum over $\pi \in \Pi$. □

So, Lemma 7.6.3 provides uniform convergence of v_n^π and v_n^* as $n \to \infty$. In analogy to v_∞^π, define $v_\infty^* := \lim_{n\to\infty} v_n^*$. Since v_n^* is increasing by Lemma 7.6.1, the limit is well-defined. Furthermore, we have $v_\infty^* \geq v_n^*$ for all $n \in \mathbb{N}_0$. Analogously to v_∞^π, we can identify v_∞^* with v^*.

Proposition 7.6.4. *It holds*

$$v_\infty^* = v^*.$$

Proof. Let $n \in \mathbb{N}_0$. Then we have $v_n^* \leq v^* \leq v_n^* + U\beta^{n+1}/(1-\beta)$ in view of Lemmas 7.6.1 and 7.6.3. Hence $v_n^* \to v^*$ as $n \to \infty$. □

From the last assertions, we can immediately give a value iteration algorithm to determine v^*.

Proposition 7.6.5. *Let $s \in S$, $\xi \in \mathbb{R}$ and $v_{-1}^*(s,\xi) := \xi^-$, $s \in S$, $\xi \in \mathbb{R}$. Then*

$$v^* = \lim_{n\to\infty} T^n v_{-1}^*.$$

Proof. The statement follows from Proposition 7.5.4 and Proposition 7.6.4 since for $n \in \mathbb{N}$

$$T^n v_{-1}^*(s,\xi) \overset{\text{Proposition 7.5.4}}{=} v_{n-1}^*(s,\xi) \overset{\text{Proposition 7.6.4}}{\to} v^*(s,\xi), \quad n \to \infty.$$ □

Now, we have everything in hand to prove the existence of a deterministic ε-optimal policy.

Definition 7.6.6. Let $\varepsilon > 0$. For a fixed $\xi \in \mathbb{R}$, a policy $\pi \in \Pi$ is called *ε-optimal,* if $v^\pi(x_0,\xi) \leq v^*(x_0,\xi) + \varepsilon$ for all $x_0 \in S$.

Of course, for every $\varepsilon > 0$ an ε-optimal policy exists by definition. The next proposition shows that there is a deterministic ε-optimal policy, which is constructed by truncating the costs after a certain time step.

Proposition 7.6.7. *For every $\varepsilon > 0$, there exists a deterministic ε-optimal policy for problem* (7.11).

Proof. Let $\xi \in \mathbb{R}$. There exists some $n_0(\varepsilon) \in \mathbb{N}_0$ such that $U\beta^{n_0+1}/(1-\beta) < \varepsilon/2$. Let $\pi_\xi^{n_0,*} \in \Pi^d$ be optimal for the n_0-horizon problem as in the proof of Proposition 7.5.4. From Lemma 7.6.3, we obtain

$$\begin{aligned}\left|v^*(s,\xi) - v^{\pi_\xi^{n_0,*}}(s,\xi)\right| &\leq \left|v^*(s,\xi) - v_{n_0}^*(s,\xi)\right| + \left|v_{n_0}^*(s,\xi) - v^{\pi_\xi^{n_0,*}}(s,\xi)\right| \\ &= \left|v^*(s,\xi) - v_{n_0}^*(s,\xi)\right| + \left|v_{n_0}^{\pi_\xi^{n_0,*}}(s,\xi) - v^{\pi_\xi^{n_0,*}}(s,\xi)\right| \overset{\text{Lemma 7.6.3}}{\leq} \frac{U\beta^{n_0+1}}{1-\beta} + \frac{U\beta^{n_0+1}}{1-\beta} < \varepsilon\end{aligned}$$

for all $s \in S$. Hence, $\pi_\xi^{n_0,*}$ is ε-optimal. □

Remark 7.6.8.

1. When there are negative costs included in W, i. e., W is an arbitrary non-empty finite subset of $\mathbb{R}$, the results from this section are true, too. To see this, let $L := \min W \leq 0$. The second assertion of Lemma 7.6.3 can be modified to $v_n^\pi + L\beta^{n+1}/(1-\beta) \leq v^\pi \leq v_n^\pi + U\beta^{n+1}/(1-\beta)$ and $v_n^* + L\beta^{n+1}/(1-\beta) \leq v^* \leq v_n^* + U\beta^{n+1}/(1-\beta)$, $n \in \mathbb{N}_0$, from which it directly follows that $v_n^\pi \to v^\pi$ for every $\pi \in \Pi$ and that $v_n^* \to v^*$ as $n \to \infty$, which are the analogous results of Propositions 7.6.2 and 7.6.4. The remaining results can be established as above.

2. Also, when terminal costs are included in the finite-horizon problems, the results of this section can be established. In the analysis, the terminal costs uniformly tend to zero as $n \to \infty$, which achieves the corresponding results of this section.

The Optimality Equation

In this section, we make another approach to solve problem (7.11). In the end of this section, we are able to prove the existence of a deterministic optimal policy of the intermediate problem (7.11). Furthermore, we construct an optimal policy. The results are based on the Banach fixed point theorem. First, we give a lemma similar to Lemma 7.5.1.

Lemma 7.6.9. *Let $\pi \in \Pi$. Then it holds*

$$v^\pi(s,\xi) = \beta \sum_{a_0 \in D(s)} \pi_0(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} \, v^{{}^1\pi^{(s,a_0,c_0)}}\left(s', \frac{\xi - c_0}{\beta}\right), \quad s \in S,\ \xi \in \mathbb{R}.$$

Proof. For $s \in S$, $\xi \in \mathbb{R}$ and $\pi \in \Pi$, one computes

$$\begin{aligned}v^\pi(s,\xi) &= E^\pi\left[(C^\infty - \xi)^+ \,\middle|\, X_0 = s\right] \\ &= \sum_{a_0 \in D(s)} \pi(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} E^\pi\left[\left(c_0 + \beta C_1 + \beta^2 C_2 + \ldots - \xi\right)^+ \,\middle|\, X_0 = s, A_0 = a_0, C_0 = c_0, X_1 = s'\right] \\ &= \beta \sum_{a_0 \in D(s)} \pi(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} E^\pi\left[\left(C_1 + \beta C_2 + \ldots - \frac{\xi - c_0}{\beta}\right)^+ \,\middle|\, X_0 = s, A_0 = a_0, C_0 = c_0, X_1 = s'\right] \\ &= \beta \sum_{a_0 \in D(s)} \pi(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} E^{{}^1\pi^{(s,a_0,c_0)}}\left[\left(C_0 + \beta C_1 + \ldots - \frac{\xi - c_0}{\beta}\right)^+ \,\middle|\, X_0 = s'\right] \\ &= \beta \sum_{a_0 \in D(s)} \pi(a_0|s) \sum_{\substack{c_0 \in W,\\ s' \in S}} p_{ss'c_0}^{a_0} \, v^{{}^1\pi^{(s,a_0,c_0)}}\left(s', \frac{\xi - c_0}{\beta}\right),\end{aligned}$$

which is the assertion. □

A short formulation of Lemma 7.6.9 shall be

$$v^{\pi} = T_{\pi_0} v^{{}^1\pi}.$$

The result of Lemma 7.6.9 can be inductively expanded to $l \in \mathbb{N}$ time steps. Therefore, let for every $l \in \mathbb{N}$

$$\begin{aligned} {}^l\pi_k^{(x_0,a_0,c_0,\dots,x_{l-1},a_{l-1},c_{l-1})}(a|h_k) &:= \pi_{k+l}(a|x_0,a_0,c_0,\dots,x_{l-1},a_{l-1},c_{l-1},h_k), \\ &\quad x_m \in S,\, a_m \in D(x_m),\, c_m \in W,\, m = 0,\dots,l-1,\, h_k \in H_k,\, a \in A,\, k = 0,1,\dots, \end{aligned}$$

be the *l-cut-head policy of* π for every $\pi \in \Pi$. With this definition, we can formulate the following corollary of Lemma 7.6.9.

Corollary 7.6.10. *Let $\pi \in \Pi$. Then for every $l \in \mathbb{N}$*

$$\begin{aligned} v^{\pi}(x_0,\xi) = \beta^l &\sum_{a_0 \in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0 \in W,\\ x_1 \in S}} p^{a_0}_{x_0 x_1 c_0} \sum_{a_1 \in D(x_1)} {}^1\pi_0^{(x_0,a_0,c_0)}(a_1|x_1) \sum_{\substack{c_1 \in W,\\ x_2 \in S}} p^{a_1}_{x_1 x_2 c_1} \cdot \dots \\ &\cdot \sum_{a_l \in D(x_l)} {}^l\pi_0^{(x_0,a_0,c_0,\dots,x_{l-1},a_{l-1},c_{l-1})}(a_l|x_l) \\ &\cdot \sum_{\substack{c_l \in W,\\ x_{l+1} \in S}} p^{a_l}_{x_l x_{l+1} c_l} v^{{}^{l+1}\pi(x_0,a_0,c_0,\dots,x_l,a_l,c_l)}\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(x - \sum_{k=0}^{l} \beta^k c_k\right)\right), \quad x_0 \in S,\, \xi \in \mathbb{R}. \end{aligned} \tag{7.12}$$

Proof. Let $\pi \in \Pi$, $x_0 \in S$ and $\xi \in \mathbb{R}$. For $l = 1$, the assertion is proven in Lemma 7.6.9. Let the assertion be true for some $l \in \mathbb{N}$. Similar to the proof of Lemma 7.6.9, we obtain

$$\begin{aligned} &v^{{}^{l+1}\pi(x_0,a_0,c_0,\dots,x_l,a_l,c_l)}\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right) \\ &\quad = \beta \sum_{a_{l+1} \in D(x_{l+1})} {}^{l+1}\pi_0^{(x_0,a_0,c_0,\dots,x_l,a_l,c_l)}(a_{l+1}|x_{l+1}) \\ &\qquad \cdot \sum_{\substack{c_{l+1} \in W,\\ x_{l+2} \in S}} p^{a_{l+1}}_{x_{l+1} x_{l+2} c_{l+1}} v^{{}^{l+2}\pi(x_0,a_0,c_0,\dots,x_{l+1},a_{l+1},c_{l+1})}\left(x_{l+2}, \frac{1}{\beta^{l+2}}\left(\xi - \sum_{k=0}^{l+1} \beta^k c_k\right)\right) \end{aligned} \tag{7.13}$$

for every $x_{l+1}, x_m \in S$, $a_m \in D(x_m)$, $c_m \in W$, $m = 0,\dots,l$. Plugging in (7.13) into the induction hypothesis (7.12), yields the assertion. □

In what follows next, we want to show a little more than we have established up to this point. We show the convergence of $T^n v^*_{-1} \overset{n\to\infty}{\to} v^*$ for a broader class of functions, i. e., $T^n u \overset{n\to\infty}{\to} v^*$ for all u lying in the set of functions $\mathscr{F}$, which is defined by

$$\begin{aligned} \mathscr{F} := \Bigg\{ u : S \times \mathbb{R} \to \mathbb{R}_{\geq 0} \;\Bigg|\; & u(s,\cdot) \text{ decreasing } (s \in S),\ |u(s,\xi) - u(s,\eta)| \leq |\xi - \eta| \ (s \in S,\, \xi,\, \eta \in \mathbb{R}), \\ & \frac{\partial}{\partial \xi} u(s,\xi) = -1 \ (s \in S,\, \xi < 0),\ u(s,\xi) = 0 \ (s \in S,\, \xi \geq U/(1-\beta)) \Bigg\}. \end{aligned}$$

Note that every function $u \in \mathscr{F}$ is unbounded since $u(s,\xi) \overset{\xi\to-\infty}{\to} \infty$ for all $s \in S$. Now, we derive that $v^{\pi} \in \mathscr{F}$ for all $\pi \in \Pi$ and that $(\mathscr{F}, d)$ is a complete metric space, where for u, $v \in \mathscr{F}$, the metric d is defined by

$$d(u,v) := \sup_{\substack{s \in S,\\ \xi \in \mathbb{R}}} |u(s,\xi) - v(s,\xi)|.$$

Note that d is indeed a metric on $\mathscr{F}$: for u, $v \in \mathscr{F}$, it holds $0 \leq d(u,v) \leq U/(1-\beta) < \infty$ since $u(s,\cdot)$ and $v(s,\cdot)$ are continuous, equal zero above $U/(1-\beta)$, since for each $s \in S$ the difference $u(s,\xi) - v(s,\xi)$ is constant for all $\xi \leq 0$ and since S is finite. Further, it holds $u = v$ if and only if $d(u,v) = 0$. Symmetry and triangle inequality are obvious. In order to apply the Banach fixed point theorem, we show that the operator T maps $\mathscr{F}$ into $\mathscr{F}$ itself and that T is a contraction mapping. But first, we illustrate the reason of defining $\mathscr{F}$ in the above manner.

Proposition 7.6.11. *Let $\pi \in \Pi$, then $v^\pi \in \mathscr{F}$.*

Proof. Let $\pi \in \Pi$. At first, $v^\pi \geq 0$. Since $(\cdot)^+$ is increasing, $v^\pi(s,\cdot)$ is decreasing for all $s \in S$. The Lipschitz continuity of $v^\pi(s,\cdot)$ can be achieved as in the proof of Proposition 7.5.9. For $s \in S$ and $\xi < 0$, we have

$$v^\pi(s,\xi) = E^\pi\left[(C^\infty - \xi)^+ \,\middle|\, X_0 = s\right] = E^\pi\left[C^\infty \,|\, X_0 = s\right] - \xi$$

since $C^\infty \geq 0$ P-a. s., and hence $\partial v^\pi(s,\xi)/\partial\xi = -1$ for all $s \in S$ and for all $\xi < 0$. Since $C^\infty \leq U/(1-\beta)$ P^π-a. s., it holds $v^\pi(s,\xi) = 0$ for all $s \in S$ and for all $\xi \geq U/(1-\beta)$. So, v^π satisfies all conditions of functions lying in $\mathscr{F}$, and therefore, $v^\pi \in \mathscr{F}$. □

Lemma 7.6.12. *The metric space $(\mathscr{F}, d)$ is complete.*

Proof. We have to show that every Cauchy sequence in $\mathscr{F}$ converges towards an element of $\mathscr{F}$ with respect to the metric d. To this end, let $(u_n)_{n\in\mathbb{N}_0}$ be a Cauchy sequence in $\mathscr{F}$. Let $\varepsilon > 0$. Then there exists some $n_0(\varepsilon) \in \mathbb{N}_0$ such that $d(u_m, u_n) < \varepsilon$ for all $m, n \geq n_0$. Let $s \in S$ and $\xi \in \mathbb{R}$. Then $|u_m(s,\xi) - u_n(s,\xi)| \leq d(u_m,u_n) < \varepsilon$ for all $m, n \geq n_0$. Hence, $(u_n(s,\xi))_{n\in\mathbb{N}_0}$ is a Cauchy sequence in $\mathbb{R}$. Let $u(s,\xi) \coloneqq \lim_{n\to\infty} u_n(s,\xi)$, which exists for all $s \in S$ and $\xi \in \mathbb{R}$ since $(\mathbb{R}, |\cdot|)$ is complete. In this manner, define a function $u : S \times \mathbb{R} \to \mathbb{R}$ by $u(s,\xi) \coloneqq \lim_{n\to\infty} u_n(s,\xi)$. The definition of u is independent of the choice of the Cauchy sequence $(u_n)_{n\in\mathbb{N}_0}$. We have to show that $u \in \mathscr{F}$ and $u_n \overset{n\to\infty}{\to} u$ with respect to the metric d.

1. Since $u_n(s,\cdot)$ is decreasing for all $s \in S$, we have $u(s,\cdot)$ is decreasing by Lemma A.7.
2. The Lipschitz continuity of $u(s,\cdot)$ with Lipschitz constant 1 can be seen as follows: let $\xi, \eta \in \mathbb{R}$, then
$$|u(s,\xi) - u(s,\eta)| \leq |u(s,\xi) - u_n(s,\xi)| + |u_n(s,\xi) - u_n(s,\eta)| + |u_n(s,\eta) - u(s,\eta)| \leq \varepsilon + |\xi - \eta|$$
for all $n \geq n_0$ with $n_0 \in \mathbb{N}_0$ such that $|u(s,\xi) - u_n(s,\xi)| < \varepsilon/2$ and $|u_n(s,\eta) - u(s,\eta)| < \varepsilon/2$ for all $n \geq n_0$.
3. Let $s \in S$. For every $\xi < 0$, we have that $u_n = a_n - \xi$ for certain $a_n \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_0$. Thus $u(s,\xi) = \lim_{n\to\infty} u_n(s,\xi) = \lim_{n\to\infty} a_n - \xi$ and $\partial u(s,\xi)/\partial\xi = -1$.
4. Since $u_n(s,\xi) = 0$ for all $s \in S$ and for all $\xi > U/(1-\beta)$, we have $u(s,\xi) = 0$ for all $s \in S$ and for all $\xi > U/(1-\beta)$.

To show the convergence of u_n towards u with respect to the metric d, let $\varepsilon > 0$ and $n_0(\varepsilon) \in \mathbb{N}_0$ such that $d(u_m,u_n) < \varepsilon$ for all $m, n \geq n_0$. Then we have $|u_m(s,\xi) - u_n(s,\xi)| \leq d(u_m,u_n) < \varepsilon$ for all $s \in S$, $\xi \in \mathbb{R}$ and $m, n \geq n_0$. Letting $m \to \infty$, this implies $|u(s,\xi) - u_n(s,\xi)| < \varepsilon$ for all $s \in S$, $\xi \in \mathbb{R}$ and $n \geq n_0$, and hence $d(u,u_n) < \varepsilon$ for all $n \geq n_0$. □

Lemma 7.6.13. *The operator T is a contraction mapping with Lipschitz constant β on the function space containing all functions $u : S \times \mathbb{R} \to \mathbb{R}$ with respect to the metric d.*

Proof. Let $u, v : S \times \mathbb{R} \to \mathbb{R}$ with $d(u,v) < \infty$. Then Tu and Tv are both functions with domain $S \times \mathbb{R}$ and values in $\mathbb{R}$. Now, we check the contraction property: let $s \in S$ and $\xi \in \mathbb{R}$, then

$$\begin{aligned}|Tu(s,\xi) - Tv(s,\xi)| &\overset{\text{Lemma A.5}}{\leq} \beta \max_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{ss'c_0} \left|u\left(s', \frac{\xi - c_0}{\beta}\right) - v\left(s', \frac{\xi - c_0}{\beta}\right)\right| \leq \beta \max_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{ss'c_0}\, d(u,v) \\ &= \beta\, d(u,v).\end{aligned}$$

Taking the supremum on the left hand side over all $s \in S$ and $\xi \in \mathbb{R}$ yields the assertion. □

Lemma 7.6.14. *The operator T maps $\mathscr{F}$ into itself, i. e., $T\mathscr{F} \subset \mathscr{F}$.*

Proof. Let $u \in \mathscr{F}$. Then $Tu \geq 0$. We have to show that Tu satisfies the four remaining properties of functions belonging to $\mathscr{F}$:

1. The functions $Tu(s,\cdot)$ are decreasing since the functions $u(s,\cdot)$ are decreasing for all $s \in S$.
2. For $\xi, \eta \in \mathbb{R}$, we have
$$\begin{aligned}|Tu(s,\xi) - Tu(s,\eta)| &\overset{\text{Lemma A.5}}{\leq} \beta \max_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{ss'c_0} \left|u\left(s', \frac{\xi - c_0}{\beta}\right) - u\left(s', \frac{\eta - c_0}{\beta}\right)\right| \\ &\leq \beta \max_{a\in D(s)} \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{ss'c_0} \left|\frac{\xi - c_0}{\beta} - \frac{\eta - c_0}{\beta}\right| = |\xi - \eta|,\end{aligned}$$
since $u(s,\cdot)$ is Lipschitz continuous with Lipschitz constant 1 for all $s \in S$.

3. Let $\xi < 0$. Then $(\xi - c_0)/\beta < 0$ for all $c_0 \in W$. Since $u \in \mathscr{F}$, we can write $u(s,\xi) \coloneqq a_s - \xi$, $s \in S$, $\xi < 0$, for certain constants $a_s \geq 0$, $s \in S$. Then we have for $\xi, \eta < 0$

$$\begin{aligned} Tu(s,\xi) - Tu(s,\eta) &= \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} u\left(s', \frac{\xi - c_0}{\beta}\right) - \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} u\left(s', \frac{\eta - c_0}{\beta}\right) \\ &= \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} \left(a_{s'} - \frac{\xi - c_0}{\beta}\right) - \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} \left(a_{s'} - \frac{\eta - c_0}{\beta}\right) \\ &= -\xi + \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} \left(a_{s'} + \frac{c_0}{\beta}\right) + \eta - \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} \left(a_{s'} + \frac{c_0}{\beta}\right) = \eta - \xi. \end{aligned}$$

Hence, $\partial Tu(s,\xi)/\partial \xi = -1$ for all $s \in S$ and for all $\xi < 0$.

4. For every $c_0 \in W$, it holds

$$\begin{aligned} &U - c_0 \geq 0 \quad \Leftrightarrow \quad (1-\beta)(U - c_0) \geq 0 \quad \Leftrightarrow \quad U - (1-\beta)c_0 \geq \beta U \quad \Leftrightarrow \quad \frac{U}{(1-\beta)\beta} - \frac{c_0}{\beta} \geq \frac{U}{1-\beta} \\ &\Leftrightarrow \quad \frac{\frac{U}{1-\beta} - c_0}{\beta} \geq \frac{U}{1-\beta}. \end{aligned}$$

Since $u(s,\cdot)$ is decreasing for $s \in S$, we have for $s' \in S$, $\xi \geq U/(1-\beta)$ and $c_0 \in W$

$$0 \leq u\left(s', \frac{\xi - c_0}{\beta}\right) \leq u\left(s', \frac{\frac{U}{1-\beta} - c_0}{\beta}\right) \leq u\left(s', \frac{U}{1-\beta}\right) = 0,$$

from which we conclude

$$Tu(s,\xi) = \beta \min_{a \in D(s)} \sum_{\substack{c_0 \in W, \\ s' \in S}} p^a_{ss'c_0} u\left(s', \frac{\xi - c_0}{\beta}\right) = 0$$

for all $\xi \geq U/(1-\beta)$.

From 1–4, we have $Tu \in \mathscr{F}$. □

In Lemmas 7.6.12–7.6.14, we have gathered all conditions to apply the Banach fixed point theorem:

Proposition 7.6.15. *The optimal value function v^* is the unique fixed point of the operator T in the function space $\mathscr{F}$. Furthermore, we have $T^n u \overset{n\to\infty}{\to} v^*$ for all $u \in \mathscr{F}$.*

Proof. By Lemmas 7.6.13 and 7.6.14, T is a contraction mapping such that $T\mathscr{F} \subset \mathscr{F}$. By the Banach fixed point theorem (cf. (Agarwal et al., 2009), Theorem 4.1.5), we have $T^n v \overset{n\to\infty}{\to} v_\infty$ for all $v \in \mathscr{F}$ for some unique $v_\infty \in \mathscr{F}$. On the other hand, since $T^n v^*_{-1} \overset{n\to\infty}{\to} v^*$ and $v^*_{-1} \in \mathscr{F}$ by Proposition 7.6.5, we conclude $v_\infty = v^*$. □

Remark 7.6.16.

1. By Remark 7.5.5.1, an MDP with state space $S \times \mathbb{R}$ is defined by the operator T where the decision maker has to pay a cost of zero at each time step. The expected total discounted cost is then zero for every policy. Hence, the optimal expected total discounted cost is zero for all initial states. Indeed, $v \equiv 0$ is a fixed point of T. Since the costs are bounded, v is the unique fixed point of T in the space of all real-valued bounded measurable functions by Note 4.2.1 of (Hernández-Lerma and Lasserre, 1996). But $0 \notin \mathscr{F}$. Therefore, the operator T might have several fixed points. But there is only one fixed point in the set $\mathscr{F}$ by Proposition 7.6.15.

2. The inner problem (7.6) can be formulated as an MDP with state space $S \times \mathbb{R}$ with terminal cost $v^*_{-1}(s,\xi) = \xi^-$, $(s,\xi) \in S \times \mathbb{R}$, in the same manner as in Remark 7.5.5.2, we can apply theory of (Bertsekas and Shreve, 1978; Schäl, 2004) since the respective assumptions, i. e., the uniform increase assumption and the structure assumption, are satisfied. Therefore, the optimality equation as well as the value iteration hold. But since we could use the Banach fixed point theorem, we were able show a little more than can be established with the general theory, e. g., the convergence $T^n u \to v^*$ for every $u \in \mathscr{F}$.

In order to provide a deterministic optimal policy for the infinite-horizon problem, we show that the value of every history-dependent randomized policy can be generated from any function $u \in \mathscr{F}$ by inductively applying the appropriate l-step iteration.

Proposition 7.6.17. *Let $\pi \in \Pi$ and $u \in \mathscr{F}$. Further, let*

$$u_{l+1}(x_0,\xi) := \beta^l \sum_{a_0 \in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0 \in W,\\ x_1 \in S}} p^{a_0}_{x_0 x_1 c_0} \sum_{a_1 \in D(x_1)} {}^1\pi_0^{(x_0,a_0,c_0)}(a_1|x_1) \sum_{\substack{c_1 \in W,\\ x_2 \in S}} p^{a_1}_{x_1 x_2 c_1} \cdot \ldots$$
$$\cdot \sum_{a_l \in D(x_l)} {}^l\pi_0^{(x_0,a_0,c_0,\ldots,x_{l-1},a_{l-1},c_{l-1})}(a_l|x_l) \sum_{\substack{c_l \in W,\\ x_{l+1} \in S}} p^{a_l}_{x_l x_{l+1} c_l}\, u\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right)$$

for $l \in \mathbb{N}_0$, $x_0 \in S$, $\xi \in \mathbb{R}$. Then

$$d\,(u_l, v^\pi) \to 0, \quad l \to \infty.$$

Proof. Let $l \in \mathbb{N}$. By Corollary 7.6.10, we have

$$\begin{aligned}
&|v^\pi(x_0,\xi) - u_l(x_0,\xi)| \\
&= \Bigg| \beta^l \sum_{a_0 \in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0 \in W,\\ x_1 \in S}} p^{a_0}_{x_0 x_1 c_0} \sum_{a_1 \in D(x_1)} {}^1\pi_0^{(x_0,a_0,c_0)}(a_1|x_1) \sum_{\substack{c_1 \in W,\\ x_2 \in S}} p^{a_1}_{x_1 x_2 c_1} \cdot \ldots \\
&\quad \cdot \sum_{a_l \in D(x_l)} {}^l\pi_0^{(x_0,a_0,c_0,\ldots,x_{l-1},a_{l-1},c_{l-1})}(a_l|x_l) \\
&\quad\quad \cdot \sum_{\substack{c_l \in W,\\ x_{l+1} \in S}} p^{a_l}_{x_l x_{l+1} c_l}\, v^{{}^{l+1}\pi^{(x_0,a_0,c_0,\ldots,x_l,a_l,c_l)}}\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right) \\
&\quad - \beta^l \sum_{a_0 \in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0 \in W,\\ x_1 \in S}} p^{a_0}_{x_0 x_1 c_0} \sum_{a_1 \in D(x_1)} {}^1\pi_0^{(x_0,a_0,c_0)}(a_1|x_1) \sum_{\substack{c_1 \in W,\\ x_2 \in S}} p^{a_1}_{x_1 x_2 c_1} \cdot \ldots \\
&\quad\quad \cdot \sum_{a_l \in D(x_l)} {}^l\pi_0^{(x_0,a_0,c_0,\ldots,x_{l-1},a_{l-1},c_{l-1})}(a_l|x_l) \sum_{\substack{c_l \in W,\\ x_{l+1} \in S}} p^{a_l}_{x_l x_{l+1} c_l}\, u\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right) \Bigg| \\
&= \Bigg| \beta^l \sum_{a_0 \in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0 \in W,\\ x_1 \in S}} p^{a_0}_{x_0 x_1 c_0} \sum_{a_1 \in D(x_1)} {}^1\pi_0^{(x_0,a_0,c_0)}(a_1|x_1) \sum_{\substack{c_1 \in W,\\ x_2 \in S}} p^{a_1}_{x_1 x_2 c_1} \cdot \ldots \\
&\quad \cdot \sum_{a_l \in D(x_l)} {}^l\pi_0^{(x_0,a_0,c_0,\ldots,x_{l-1},a_{l-1},c_{l-1})}(a_l|x_l) \sum_{\substack{c_l \in W,\\ x_{l+1} \in S}} p^{a_l}_{x_l x_{l+1} c_l} \\
&\quad\quad \cdot \left[v^{{}^{l+1}\pi^{(x_0,a_0,c_0,\ldots,x_l,a_l,c_l)}}\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right) - u\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right)\right] \Bigg| \\
&\le \beta^l \sum_{a_0 \in D(s)} \pi_0(a_0|x_0) \sum_{\substack{c_0 \in W,\\ x_1 \in S}} p^{a_0}_{x_0 x_1 c_0} \sum_{a_1 \in D(x_1)} {}^1\pi_0^{(x_0,a_0,c_0)}(a_1|x_1) \sum_{\substack{c_1 \in W,\\ x_2 \in S}} p^{a_1}_{x_1 x_2 c_1} \cdot \ldots \\
&\quad \cdot \sum_{a_l \in D(x_l)} {}^l\pi_0^{(x_0,a_0,c_0,\ldots,x_{l-1},a_{l-1},c_{l-1})}(a_l|x_l) \sum_{\substack{c_l \in W,\\ x_{l+1} \in S}} p^{a_l}_{x_l x_{l+1} c_l} \\
&\quad\quad \cdot \underbrace{\left| v^{{}^{l+1}\pi^{(x_0,a_0,c_0,\ldots,x_l,a_l,c_l)}}\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right) - u\left(x_{l+1}, \frac{1}{\beta^{l+1}}\left(\xi - \sum_{k=0}^{l} \beta^k c_k\right)\right)\right|}_{\le \frac{U}{1-\beta}}
\end{aligned}$$

$$\leq \frac{\beta^l U}{1-\beta}$$

since $v^{l+1\pi^{(x_0,a_0,c_0,\ldots,x_l,a_l,c_l)}} \in \mathscr{F}$ for every $x_m \in S$, $a_m \in D(x_m)$, $c_m \in W$, $m = 0, 1, \ldots, l$, by Proposition 7.6.11. Thus $d(u_l, v^\pi) \leq \beta^l U/(1-\beta) \overset{l\to\infty}{\to} 0$. □

At last, we are able to prove that there exists a deterministic optimal policy for the infinite-horizon problem.

Proposition 7.6.18. *There exists a deterministic optimal policy $\pi^*_\xi \in \Pi^d$ for problem* (7.11)*.*

Proof. Inductively define the following deterministic history-dependent policy $\pi^*_\xi \in \Pi^d$ for $\xi \in \mathbb{R}$:

$$\begin{aligned}
&\left(\pi^*_\xi\right)_0(x_0) := \mu^*_\xi(x_0) \quad \text{such that} \quad \mu^*_\xi(x_0) \in \operatorname*{arg\,min}_{a\in D(x_0)} \left\{ \sum_{\substack{c_0\in W,\\ s'\in S}} p^a_{x_0 s' c_0}\, v^*\left(s', \frac{\xi - c_0}{\beta}\right) \right\},\\
&\left(\pi^*_\xi\right)_1(c_0, x_1) := \left(\pi^*_\xi\right)_1(x_0, a_0, c_0, x_1) := \left(\pi^*_{\frac{\xi-c_0}{\beta}}\right)_0(x_1),\\
&\left(\pi^*_\xi\right)_k(c_0, c_1, \ldots, c_{k-1}, x_k) := \left(\pi^*_\xi\right)_k(x_0, a_0, c_0, x_1, \ldots, x_{k-1}, a_{k-1}, c_{k-1}, x_k) := \left(\pi^*_{\frac{\xi-c_0}{\beta}}\right)_{k-1}(c_1, \ldots, c_{k-1}, x_k),\\
&\quad = \mu^*_{\frac{1}{\beta^k}\left(\xi - \sum_{l=0}^{k-1}\beta^l c_l\right)}(x_k),\\
&\qquad\qquad x_i, x_k \in S,\ a_i \in D(x_i),\ c_i \in W,\ i = 0, 1, \ldots, k-1,\ k = 2, 3, \ldots.
\end{aligned} \tag{7.14}$$

Since $v^* = Tv^* = T^2v^* = \ldots$ by Proposition 7.6.15, we have for every $l \in \mathbb{N}$, $x_0 \in S$ and $\xi \in \mathbb{R}$ by the definition of π^*_ξ

$$\begin{aligned}
v^*(x_0, \xi) = T^l v^*(x_0, \xi) = \beta^l \sum_{\substack{c_0\in W,\\ x_1\in S}} p^{\left(\pi^*_\xi\right)_0(x_0)}_{x_0 x_1 c_0} \sum_{\substack{c_1\in W,\\ x_2\in S}} p^{{}^1\left(\pi^*_\xi\right)^{(x_0,a_0,c_0)}_0(x_1)}_{x_1 x_2 c_1} \cdot \ldots \cdot \sum_{\substack{c_l\in W,\\ x_{l+1}\in S}} p^{{}^l\left(\pi^*_\xi\right)^{(x_0,a_0,c_0,\ldots,x_{l-1},a_{l-1},c_{l-1})}_0(x_l)}_{x_l x_{l+1} c_l}\\
\cdot v^*\left(x_{l+1}, \frac{1}{\beta^l}\left(\xi - \sum_{m=0}^{l-1}\beta^m c_m\right)\right)\\
\to v^{\pi^*_\xi}(x_0, \xi), \quad l \to \infty,
\end{aligned}$$

where the convergence is provided by Proposition 7.6.17 since $v^* \in \mathscr{F}$ by Proposition 7.6.15. Hence $v^*(x_0, \xi) = v^{\pi^*_\xi}(x_0, \xi)$, or in other words, π^*_ξ is optimal for problem (7.11). □

Remark 7.6.19.

1. Note that the optimal policy defined in Proposition 7.6.18 does not depend on the complete history by the last equation of (7.14). Similar to the finite-horizon problem, it is sufficient to know the accumulated discounted cost $\sum_{l=0}^k \beta^l c_l$, the current system state and the current time step k in order to derive an optimal policy.
2. The infinite-horizon problem (7.11) can also be formulated as an MDP with state space $S \times \mathbb{R}$ in the same way as in Remark 7.6.16.2. Then we see that the optimal policy given by (7.14) is stationary with respect to the state space $S \times \mathbb{R}$. This result can also be derived from Theorem 3 of (Schäl, 2004) since the so-called uniform increase assumption and the structure assumption both are satisfied.
3. As the proof shows, after each time step and resulting cost, the decision maker has to adjust her preference, which ξ' should be the new threshold for the following infinite time steps in order to act optimally corresponding to the original threshold ξ with respect to the intermediate criterion.
4. We could also have included negative costs as long as W remains finite. The function space $\mathscr{F}$ then should be changed to

$$\begin{aligned}
\mathscr{F} := \Bigg\{ u : S \times \mathbb{R} \to \mathbb{R}_{\geq 0} \;\Bigg|\; & u(s,\cdot) \text{ decreasing } (s \in S),\ |u(s,\xi) - u(s,\eta)| \leq |\xi - \eta| \ (s \in S, \xi, \eta \in \mathbb{R}),\\
& \frac{\partial}{\partial \xi} u(s,\xi) = -1 \ (s \in S, \xi < L/(1-\beta)), u(s,\xi) = 0 \ (s \in S, \xi \geq U/(1-\beta)) \Bigg\},
\end{aligned}$$

where $L := \min W$, again. The results of this section can be derived with simple modifications of the above proofs.

5. (Schäl, 1990) studies discrete-time MDPs with the non-standard infinite horizon lim sup criterion, as he calls it. At each time step, the decision maker has to pay a cost which is not discounted and after the nth time step she has to pay a terminal cost. The largest class of policies (Schäl, 1990) examines contains all deterministic policies which depend on the history of the preceding states. This class shall be denoted by Π_X. The total cost up to time step $n \in \mathbb{N}_0$ when starting in initial state X_0 is given by

$$u_n(\pi) := \sum_{k=0}^{n} c(X_k, \pi(X_0, \dots, X_k)) + u(X_n).$$

(Schäl, 1990) considers the case when letting $n \to \infty$. The general assumptions are that for every policy π and initial state x_0 the conditional expectation

$$C(\pi, x_0) := E^{\pi}\left[\liminf_{n\to\infty} u_n(\pi) \,\middle|\, X_0 = x_0\right]$$

exists and that $-\infty < \inf_{\pi\in\Pi_X} C(\pi, x_0) < \infty$. Since we want to minimize the total discounted cost, the original lim sup when maximizing, as does (Schäl, 1990), turns into a lim inf. The intermediate criterion of this section can be seen as a special case of the lim sup criterion. To see this, we have to reformulate the MDP as follows: let $\xi \in \mathbb{R}$ be fixed. Let the state space be

$$\tilde{S} := S \times \bigcup_{k=0}^{\infty} \left(\bigtimes_{l=0}^{k} W \times \bigtimes_{l=k+1}^{\infty} \{\Delta\} \right)$$

so that the state is determined by the system state and the incurred costs. Here, we assume that $\Delta \notin W$, which marks the remaining time steps, i. e., there is a component from which on all components are Δ and none of the preceding components is equal to Δ. Note that $\tilde{S}$ is countably infinite. A state transition from $\tilde{s} = (s, c_0, c_1, \dots, c_k, \Delta, \Delta, \dots)$ to $(s', c_0, c_1, \dots, c_k, c_{k+1}, \Delta, \dots)$ under action $a \in \tilde{D}(\tilde{s}) := D(s)$ occurs with probability $p^a_{ss'c_{k+1}}$. The one-step costs satisfy $c(\tilde{s}, \tilde{a}) = 0$ for all $(\tilde{s}, \tilde{a}) \in \tilde{D}$, and the terminal cost is given by $u(\tilde{s}) := \left(\sum_{k=0}^{n} \beta^k c_k - \xi\right)^+$, $\tilde{s} = (s, c_0, c_1, \dots, c_k, \Delta, \dots) \in \tilde{S}$. The initial state is $(x_0, \Delta, \dots)$. Then we have for a deterministic policy $\pi \in \Pi^d$ in the original model which depends only on the state and the occurred costs

$$v^{\pi}(x_0, \xi) = E^{\pi}\left[\left(\sum_{k=0}^{\infty} \beta^k C_k - \xi\right)^+ \,\middle|\, X_0 = x_0\right] = E^{\pi}\left[\lim_{n\to\infty}\left(\sum_{k=0}^{n} \beta^k C_k - \xi\right)^+ \,\middle|\, X_0 = x_0\right] = C(\tilde{\pi}, x_0)$$

where $\tilde{\pi}$ is the appropriate policy for the reformulated model. Hence, we can apply the results of (Schäl, 1990). Among these are that the optimality equation holds in the reformulated model. Moreover, we can establish the existence of an optimal policy for the initial state x_0 which can be derived as a minimizer of the optimality equation. This can be done by applying Satz 18.26 where the assumptions are fulfilled by Satz 18.22 and since $u(\tilde{s})$ is bounded by $U/(1-\beta) - \xi$.

The results established in this section are a bit stronger since we derived optimality of a deterministic history-dependent (in the original formulation) policy within the class of randomized history-dependent policies. Furthermore, in order to handle the average value-at-risk criterion, we need information about the optimal value function v^* as we shall see in the next section, e. g., $v^* \in \mathscr{F}$.

7.6.2. The Average Value-at-Risk Criterion

Now, the solution of the original problem, that is to minimize the average value-at-risk at confidence level $\tau \in (0,1)$ of the total discounted cost over an infinite horizon if we start in state $x_0 \in S$, is within reach. This problem is equivalent to:

$$\text{Minimize } \xi + \frac{1}{1-\tau} v^*(x_0, \xi) \text{ over all } \xi \in \mathbb{R}. \tag{7.15}$$

To show the existence of an $\xi^* \in \mathbb{R}$ which is optimal for problem (7.15), we define similarly to the finite-horizon case the objective function of problem (7.15) by

$$w^{\tau}(x_0, \xi) := \xi + \frac{1}{1-\tau} v^*(x_0, \xi), \quad x_0 \in S, \xi \in \mathbb{R}.$$

The optimal value function v^* is an element of $\mathscr{F}$, which follows from the Banach fixed point theorem in Proposition 7.6.15. So, we have the following proposition.

Proposition 7.6.20. *There exists a solution to problem* (7.15).

Proof. Let $x_0 \in S$. It holds $w^\tau(x_0,\xi) \overset{\xi\to-\infty}{\to} \infty$, since $\partial v^*(x_0,\xi)/\partial\xi = -1$, $\xi < 0$, and $1/(1-\tau) > 1$. Moreover, we have $w^\tau(x_0,\xi) \overset{\xi\to\infty}{\to} \infty$ since $v^*(x_0,\xi) \overset{\xi\to\infty}{\to} 0$. In addition, $w^\tau(x_0,\cdot)$ is continuous since $v^*(x_0,\cdot)$ is continuous. In the same way as we established Proposition 7.5.16, we conclude the assertion. □

With the last proposition in hand, we can prove the existence of a deterministic optimal policy for problem (7.3).

Theorem 7.6.21. *There exists a deterministic optimal policy for problem* (7.3).

Proof. The proof is very similar to the finite-horizon case. Let ξ^* be a solution to problem (7.15) and $\pi^*_{\xi^*} \in \Pi^d$ be a deterministic optimal policy for the intermediate problem (7.11) with threshold ξ^* which is defined as in (7.14) for $\xi = \xi^*$. Then we have for $x_0 \in S$

$$\begin{aligned}
\inf_{\pi\in\Pi} \mathrm{AV@R}^\pi_\tau(C^\infty \mid X_0 = x_0) &\overset{\text{Theorem 6.2.3}}{=} \inf_{\pi\in\Pi}\left\{\inf_{\xi\in\mathbb{R}}\left\{\xi + \frac{1}{1-\tau}E^\pi\left[(C^\infty-\xi)^+ \,\middle|\, X_0 = x_0\right]\right\}\right\} \\
&\overset{\text{Lemma A.5}}{=} \inf_{\xi\in\mathbb{R}}\left\{\xi + \frac{1}{1-\tau}\underbrace{\inf_{\pi\in\Pi}E^\pi\left[(C^\infty-\xi)^+ \,\middle|\, X_0 = x_0\right]}_{=v^*(x_0,\xi)}\right\} \\
&\overset{\text{Proposition 7.6.20}}{=} \xi^* + \frac{1}{1-\tau}\inf_{\pi\in\Pi}E^\pi\left[(C^\infty-\xi^*)^+ \,\middle|\, X_0 = x_0\right] \\
&\overset{\text{Proposition 7.6.18}}{=} \xi^* + \frac{1}{1-\tau}E^{\pi^*_{\xi^*}}\left[(C^\infty-\xi^*)^+ \,\middle|\, X_0 = x_0\right] \geq \inf_{\xi\in\mathbb{R}}\left\{\xi + \frac{1}{1-\tau}E^{\pi^*_{\xi^*}}\left[(C^\infty-\xi)^+ \,\middle|\, X_0 = x_0\right]\right\} \\
&\overset{\text{Theorem 6.2.3}}{=} \mathrm{AV@R}^{\pi^*_{\xi^*}}_\tau(C^\infty \mid X_0 = x_0).
\end{aligned}$$

Hence, $\pi^*_{\xi^*}$ is optimal for problem (7.3). □

We conclude this section with some remarks.

Remark 7.6.22.

1. Here again, if ξ^* is the smallest solution to problem (7.15), then ξ^* is the value-at-risk at level τ of the $\mathrm{AV@R}_\tau$-optimal policy $\pi^*_{\xi^*}$ defined as in (7.14). But ξ^* need not be the optimal value-at-risk at level τ.

2. A finite number of negative costs could be included without changing the results of this section.

3. In order to act optimally, the decision maker chooses at first the threshold ξ^*, which she should not exceed in infinite time steps with respect to the intermediate criterion, by the proof of Theorem 7.6.21. Depending on the outcome of the actual cost after each step, the decision maker has to adjust her preference, which $\xi \in \mathbb{R}$ she should not exceed in the remaining infinite steps according to (7.14).

Next, we give a condition for a policy $\pi \in \Pi$ under which π is ε-optimal with respect to the $\mathrm{AV@R}_\tau$-criterion.

Proposition 7.6.23. *Let $\tau \in (0,1)$ and $\varepsilon > 0$. Let $\pi \in \Pi$ be a policy such that $\|v^\pi - v^*\|_\infty \leq (1-\tau)\,\varepsilon$. Then π is ε-optimal for the* $\mathrm{AV@R}_\tau$*-criterion.*

Proof. For every $x_0 \in S$, we compute

$$\left|\mathrm{AV@R}^\pi_\tau(C^\infty \mid X_0 = x_0) - \mathrm{AV@R}^*_\tau(C^\infty \mid X_0 = x_0)\right| \leq \frac{1}{1-\tau}\sup_{\xi\in\mathbb{R}}|v^\pi(x_0,\xi) - v^*(x_0,\xi)| \leq \varepsilon.$$

Hence, π is ε-optimal for the $\mathrm{AV@R}_\tau$-criterion. □

According to Proposition 7.6.23, the policy π has to be $((1-\tau)\,\varepsilon)$-optimal for the intermediate criterion for every threshold $\xi \in \mathbb{R}$ so that it is $\mathrm{AV@R}_\tau$-optimal. This is a quite strong assumption on π.

7.7. Conclusion

Having established the results for the average value-at-risk criterion for an infinite horizon, we now want to examine whether this criterion is applicable in practice. Recall that our goal is to develop a decision support system for the surveillance task of a critical infrastructure.

At first, it is very intricate to compute the value functions v^* or even ε-approximations. So, we cannot easily compute optimal or ε-optimal, according to Proposition 7.6.23, policies for the average value-at-risk criterion. A first idea that comes into mind is to use a suitable grid $\{\xi_i \mid i = 1, \ldots, n\}$ on $[0, U/(1-\beta)]$ and compute $v^*(x_0, \xi_i)$, $i = 1, \ldots, n$ (cf. (Wu and Lin, 1999; Boda and Filar, 2006)). Then the optimal value function $v^*(x_0, \cdot)$ could be approximated by a piecewise linear function $\tilde{v}^*(x_0, \cdot)$. Then $\tilde{v}^*(x_0, \cdot)$ could be used instead of $v^*(x_0, \cdot)$ in the preceding section. The result would be an approximated threshold value $\tilde{\xi}^*$ and an associated policy $\pi^*_{\tilde{\xi}^*}$ which could be used as an approximation of an average value-at-risk optimal policy.

But a much more serious drawback of using the average value-at-risk criterion in practice is the following: the average value-at-risk optimal policy as derived in the proof of Theorem 7.6.21 is not stationary with respect to the system state space S. So, if the current system state at some time step k is s, then some action a might be optimal. But at a later time step l, when the system state is again s some other action a' might be optimal. This behaviour of an optimal policy is not very intuitive so that it might cause confusion when using this policy as the basis of a decision support system.

In conclusion, although average value-at-risk is a suitable criterion for the surveillance task, it might be not a good choice to base a decision support system upon an optimal policy with respect to this criterion. This is due to the optimality of a policy which depends on the incurred costs. But when the personnel is properly instructed and accepts that there might be different optimal actions at different points in time, a policy based on the average value-at-risk criterion might be valuable.

8. Average Value-at-Risk Criterion for the Average Cost – A Non-Standard Objective Criterion for Discrete-Time Markov Decision Processes

In this chapter, we want to study the case of the average cost for the average value-at-risk criterion. Within the threat model for closed infrastructures, it means that the decision maker wants to act optimally so that the most devastating threat events are prevented in the long run. The approach is very different from the case of total discounted cost. As in the case of the expected cost criterion for the average cost (cf. (Puterman, 2005)), we need to go deeper into the structure of the MDP.

In the case of the average cost, the standard criterion is, as in the case of total discounted cost, to optimize the expected value of the costs and rewards respectively. There is an abundance of literature available for this kind of problem. A very detailed treatment can be found in (Puterman, 2005), chapters 8 and 9. The theory of the expected average cost for MDPs is much more intricate than the theory of the expected total discounted cost in MDPs since the chain structure of the MDP comes into play vastly. Results of this theory are that the optimal value function satisfies certain optimality equations and that there exist deterministic stationary optimal policies. Furthermore, algorithms for finding optimal policies are known, such as a linear programming approach and a policy improvement algorithm, which are both based on optimality equations. These methods can be refined by considering MDPs which have a certain chain structure such as unichain or communicating MDPs.

In this chapter, we consider the criterion of minimizing the average value-at-risk for finite-state finite-action MDPs. To this end, we compute the average value-at-risk for a random variable taking only a finite number of values in $\mathbb{R}$ since this is the foundation of the following. After that, we compute the average value-at-risk for so-called Markov reward processes. This is followed by a section which treats the case of unichain and weakly communicating MDPs. Then we introduce an approach how theoretically the average value-at-risk problem could be solved for general finite-state finite-action MDPs. This approach goes back to a decomposition of the state space introduced in (Ross and Varadarajan, 1991). We conclude this chapter with some remarks concerning average value-at-risk optimal policies for MDPs.

8.1. Literature Overview on Non-Standard Criteria for Discrete-Time Markov Decision Processes for the Average Cost

Alongside the expected average cost criterion, certain non-standard criteria for the average cost are studied by various authors. For example, certain percentile, i. e., quantile, performance criteria are considered in (Filar et al., 1995) for the average cost: for instance, the problem whether there exist policies which achieve a specified value of the average reward with a specified probability, the problem of maximizing the target for a specified quantile or conversely the problem of maximizing the quantile of a specified value.

In (Baykal-Gürsoy and Ross, 1992), two variability criteria are introduced: expected time-average variability and time-average expected variability. In the former criterion, the expected value of the long-run average of a function which measures variability of the actual reward from the average reward is considered. In the latter criterion, the long-run average of the variability between the actual costs and the expected average cost are minimized. The function measuring variability is assumed to be continuous. For both criteria, it turns out that there are deterministic stationary policies. In the survey paper (White, 1988), more non-standard criteria considering the average cost are introduced.

8.2. Average Value-at-Risk for Finite Random Variables

In this section, let X be a random variable on some probability space $(\Omega, \mathscr{A}, P)$ which takes the values $c_1, \ldots, c_N \in \mathbb{R}$ for some $N \in \mathbb{N}$. We assume that the values are ordered so that $c_1 < c_2 < \ldots < c_N$. Let the distribution of X be given by

$$P(X = c_i) = p_i, \quad i = 1, \ldots, N,$$

with $p_i \in (0,1)$, $i = 1, \ldots, N$. For some fixed confidence level $\tau \in (0,1)$, let $i_\tau \in \{1, \ldots, N\}$ be the unique index such that

$$\sum_{i=1}^{i_\tau} p_i \geq \tau > \sum_{i=1}^{i_\tau - 1} p_i,$$

i. e., $F_X^{-1}(\tau) = c_{i_\tau}$. From Proposition 8 of (Rockafellar and Uryasev, 2002), we have

$$\mathrm{AV@R}_\tau(X) = \frac{1}{1-\tau}\left[\left(\sum_{i=1}^{i_\tau} p_i - \tau\right) c_{i_\tau} + \left(\sum_{i=i_\tau+1}^{N} p_i c_i\right)\right].$$

By this formula, it is very easy to compute the average value-at-risk for finite random variables. From this formula, it can be seen that the probability atom present at ι_τ is split as mentioned in section 6.2.2.

8.3. Average Value-at-Risk in Markov Reward Processes

In this section, we have a look at finite-state Markov reward processes and determine the average value-at-risk of such a process for its average cost. In order to obtain the results of this section, we only need some standard results on ergodic Markov chains. At first, we define a Markov reward process according to (Puterman, 2005):

Definition 8.3.1. Let $S = \{1, \dots, N\}$ be a non-empty finite state space. For each $s \in S$, let $c_s \in \mathbb{R}$ with $c_s > 0$ representing costs and $c_s < 0$ representing rewards respectively that occur together with state s. Let $(X_k)_{k\in\mathbb{N}_0}$ be a homogeneous Markov chain on S with transition probability matrix P. The bivariate process $((X_k, c_{X_k}))_{k\in\mathbb{N}_0}$ is called a *Markov reward process.*

A Markov reward process is again a homogeneous Markov chain, now on the state space $S \times \mathbb{R}$. Let $((X_k, c_{X_k}))_{k\in\mathbb{N}_0}$ be a Markov reward process with transition probability matrix $P = (p_{ss'})_{s,s'\in S}$ for the Markov chain $(X_k)_{k\in\mathbb{N}_0}$. Furthermore, we partition the state space into $R \geq 1$ non-empty closed irreducible recurrent classes $\mathscr{R}_1, \dots, \mathscr{R}_R$ and the (possibly empty) class $\mathscr{T}$ of transient states. Define the random variables

$$C_n := \frac{1}{n+1}\sum_{k=0}^{n} c_{X_k}, \quad n = 0, 1, \dots,$$

representing the average cost up to time step n. We want to determine the average cost in the limit as n tends to infinity. For this purpose, we define the random variables

$$C_- := \liminf_{n\to\infty} C_n \quad \text{and} \quad C_+ := \limsup_{n\to\infty} C_n,$$

since it is not clear a priori that the limit of C_n exists.

To determine $\mathrm{AV@R}_\tau(C_- \,|\, X_0 = s)$ and $\mathrm{AV@R}_\tau(C_+ \,|\, X_0 = s)$, it is sufficient to know the distributions of C_- and C_+ respectively, which we compute in the following. To simplify the notation, we make the following assumption.

Assumption 8.3.2. Assume that the closed irreducible recurrent classes contain exactly one element, w. l. o. g., let $\mathscr{R}_s = \{s\}$, $s = 1, \dots, R$.

Proposition 8.3.3. *Under Assumption 8.3.2, it holds*

$$\mathrm{AV@R}_\tau(C_- \,|\, X_0 = x_0) = \mathrm{AV@R}_\tau(C_+ \,|\, X_0 = x_0) = c_{x_0}$$

for all $x_0 \in \{1, \dots, R\}$ *and* $\tau \in (0, 1)$.

Proof. For $x_0 \in \{1, \dots, R\}$, we have $P(X_k = x_0 \,|\, X_0 = x_0) = 1$ for all $k \in \mathbb{N}_0$. Hence $C_n = 1/(n+1)\sum_{k=0}^{n} c_{X_k} = c_{x_0}$ P-a. s., which transfers to the limits: $P(C_- = c_{x_0} \,|\, X_0 = x_0) = P(C_+ = c_{x_0}) = 1$, which finally leads to $\mathrm{AV@R}_\tau(C_- \,|\, X_0 = x_0) = \mathrm{AV@R}_\tau(C_+ \,|\, X_0 = x_0) = c_{x_0}$ for all $\tau \in (0, 1)$. □

Now, we are left with the case of starting in a transient state $s \in \mathscr{T}$. What we have to know are the absorption probabilities of $(X_k)_{k\in\mathbb{N}_0}$ in the recurrent states $1, \dots, R$, from which the average value-at-risk of C_- and C_+ can be computed, as we shall see. Define

$$\alpha(s, s') := P\left((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } s' \,\middle|\, X_0 = s\right), \quad s, s' \in S,$$

as the *absorption probabilities* of $(X_k)_{k\in\mathbb{N}_0}$. Since S is supposed to be finite, we have $\sum_{s'\in S} \alpha(s, s') = 1$ for all $s \in S$ by (Brémaud, 1999), Remark 5.1 of chapter 4. Of course, we have

$$\alpha(s, s') = 0 \quad (s \in S, s' \in \mathscr{T}). \tag{8.1}$$

Moreover, we have for the recurrent states

$$\alpha(s,s) = 1 \quad (s \in \{1,\dots,R\}), \tag{8.2}$$

which we implicitly used in the proof of Proposition 8.3.3. The absorption probabilities can be computed via first-step analysis (cf. (Brémaud, 1999), section 2.3.1), that is

$$\alpha(s,s') = \sum_{s''\in S} p_{ss''}\,\alpha(s'',s') \quad (s,\, s' \in S). \tag{8.3}$$

The system of equations in (8.3) has a unique solution due to the finiteness of S and the boundary conditions (8.1) and (8.2). The first-step analysis plays a crucial role later on. But next, we fully describe the conditional distribution of C_- and C_+ conditioned on the initial state $x_0 \in S$.

Theorem 8.3.4. *Under Assumption 8.3.2, it holds*

$$P(C_- = c_{s'} \,|\, X_0 = x_0) = P(C_+ = c_{s'} \,|\, X_0 = x_0) = \sum_{\substack{s''\in\{1,\dots,R\}:\\ c_{s''}=c_{s'}}} \alpha(x_0,s'')$$

for all x_0, $s' \in S$. Furthermore,

$$P\left(C_- \in \bigcup_{s'\in S}\{c_{s'}\} \,\middle|\, X_0 = x_0\right) = P\left(C_+ \in \bigcup_{s'\in S}\{c_{s'}\} \,\middle|\, X_0 = x_0\right) = 1 \quad (x_0 \in S).$$

Proof. Let $x_0 \in S$ and $s' \in S$. For $s' \in \{1,\dots,R\}$, define the stopping time

$$T_{s'} := \inf\left\{k \in \mathbb{N}_0 : X_k = s'\right\}.$$

Then by definition $P(T_{s'} < \infty \,|\, X_0 = x_0) = \alpha(x_0,s')$. With probability one, $T_{s'} < \infty$ for exactly one $s' \in \{1,\dots,R\}$ since S is finite. If $T_{s'} < \infty$, then $C_- = c_{s'}$ since $X_k = s'$, and thus $c_{X_k} = c_{s'}$, for all $k \geq n_0$ for some $n_0 \in \mathbb{N}_0$. By partitioning the event $\{C_- = c_{s'}\}$, we get

$$\begin{aligned} P(C_- = c_{s'} \,|\, X_0 = x_0) &= \sum_{s''=1}^{R} P(C_- = c_{s'}, T_{s''} < \infty \,|\, X_0 = x_0) = \sum_{s''=1}^{R} P(C_- = c_{s'}, C_- = c_{s''}, T_{s''} < \infty \,|\, X_0 = x_0) \\ &= \sum_{\substack{s''\in\{1,\dots,R\}:\\ c_{s''}=c_{s'}}} P(T_{s''} < \infty \,|\, X_0 = x_0) = \sum_{\substack{s''\in\{1,\dots,R\}:\\ c_{s''}=c_{s'}}} \alpha(x_0,s''). \end{aligned}$$

The statement that C_- cannot take any other values with probability one follows in a similar manner from

$$P\left(C_- \in \bigcup_{s'\in S}\{c_{s'}\} \,\middle|\, X_0 = x_0\right) = \sum_{s''=1}^{R} P\left(C_- \in \bigcup_{s'\in S}\{c_{s'}\},\ T_{s''} < \infty \,\middle|\, X_0 = x_0\right) = \sum_{s''=1}^{R} \alpha(x_0,s'') = 1.$$

Similar calculations lead to the statements for C_+. □

Together with the results of section 8.2, we are now able to determine $\mathrm{AV@R}_\tau(C_- \,|\, X_0 = x_0) = \mathrm{AV@R}_\tau(C_+ \,|\, X_0 = x_0)$, $x_0 \in S$, under Assumption 8.3.2.

The next step is to relax Assumption 8.3.2. From now on, we drop Assumption 8.3.2 and assume w. l. o. g. that the closed irreducible recurrent classes $\mathscr{R}_r$ consist of states $\{i_{r1},\dots,i_{rN_r}\}$, for some $N_r \in \mathbb{N}$, $r = 1,\dots,R$. Then the Markov chain is not absorbed by one special state but by one of the closed irreducible recurrent classes $\mathscr{R}_r$, $r = 1,\dots,R$. Therefore, we have to define the absorption probabilities appropriately by

$$\alpha(s,r) = P\left((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } \mathscr{R}_r \,\middle|\, X_0 = s\right), \quad s \in S,\, r \in \{1,\dots,R\}.$$

Theorem 8.3.5. *The conditional distributions of C_- and C_+ conditioned on the event $\{X_0 = x_0\}$, $x_0 \in S$, are given by*

$$P(C_- = c \,|\, X_0 = x_0) = P(C_+ = c \,|\, X_0 = x_0) = \sum_{r=1}^{R} \alpha(x_0,r)\,\delta_{c,\pi^r c^r}, \quad c \in \mathbb{R},$$

where π^r is the unique stationary distribution of the Markov chain $(X_k)_{k\in\mathbb{N}_0}$ restricted to $\mathscr{R}_r$, $c^r = (c_s)_{s\in\mathscr{R}_r}^\top$ is the cost vector restricted to $\mathscr{R}_r$, $r = 1,\dots,R$, and δ is the Kronecker delta. Furthermore, it holds

$$P\left(C_\in \bigcup_{r=1}^{R}\{\pi^r c^r\} \,\middle|\, X_0 = x_0\right) = P\left(C+ \in \bigcup_{r=1}^{R}\{\pi^r c^r\} \,\middle|\, X_0 = x_0\right) = 1 \quad (x_0 \in S).$$

Proof. Let $s \in R_r$ for some $r \in \{1,\dots,R\}$. Then we have

$$P(C_- = \pi^r c^r \mid X_0 = s) = P(C_+ = \pi^r c^r \mid X_0 = s) = 1$$

by an ergodic theorem for Markov chains (cf. (Brémaud, 1999), Theorem 4.1 of chapter 3.4). Hence, for arbitrary $x_0 \in S$ and $c \in \mathbb{R}$, it holds

$$P(C_- = c \,|\, X_0 = x_0) = \sum_{r=1}^{R} P\big((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } \mathscr{R}_r, C_- = c \,\big|\, X_0 = x_0\big) = \sum_{r=1}^{R} \alpha(x_0,r)\,\delta_{c,\pi^r c^r}$$

by partitioning the event $\{C_- = c\}$. Again, the same calculation gives the statement for C_+. The second assertion follows in a similar manner as the respective assertion in Theorem 8.3.4 due to the finiteness of S. □

Having computed the conditional distribution of C_- and C_+ conditioned on $\{X_0 = x_0\}$, we are able to determine $\mathrm{AV@R}_\tau(C_- \,|\, X_0 = x_0) = \mathrm{AV@R}_\tau(C_+ \,|\, X_0 = x_0)$, due to Theorem 8.3.5, by applying the results of section 8.2.

In the remainder of this chapter, we demonstrate how the average value-at-risk criterion could be solved for unichain and weakly communicating MDPs as well as for multichain MDPs afterwards. For the latter, we consider an approach to compute optimal policies within the class of stationary policies. At first, we give the main definitions for the following. Then we introduce a uniformization technique to convert the continuous-time model of the surveillance task into a discrete-time model. After that, the average value-at-risk criterion for the average cost is examined.

8.4. Definitions

In what follows, we define an MDP in a slightly different way than in the preceding chapter 7 where the costs are stochastic at each time step. Here, the costs are already determined by the state and the action. Therefore, we have to reformulate some definitions from chapter 7. We define an MDP $\Gamma = (S, A, D, P, \mathscr{C})$, with the following components:

- S is a finite *state space.*
- A is a finite *action space.*
- $D \subset S \times A$ is the *restriction set* which satisfies the condition $D(s) := \{a : (s,a) \in D\} \neq \emptyset$, $s \in S$.
- $\mathscr{C} = (c(s,a))_{(s,a)\in D}$ is the *cost structure* with $c(s,a) \geq 0$ for all $(s,a) \in D$.
- $P = (p_{ss}^a)_{s,s'\in S, a\in D(s)}$ are the *transition probabilities.*

As indicated in the definition of an MDP in section 7.2, these two notions are equivalent in the sense that either formulation can be converted into the other formulation.

In the following, we restrict ourselves to stationary, possibly randomized, policies. Again, we define histories and certain sets of policies of interest.

Definition 8.4.1. Define the *set of histories (up to time step k)* inductively by

$$H_0 := S,$$
$$H_{k+1} := H_k \times A \times S, \quad k = 0, 1, \dots.$$

At time step $k \in \mathbb{N}_0$, the decision maker has access to the information given in $h_k \in H_k$. Based on this information she is able to choose an action, possibly in a randomized manner.

Definition 8.4.2. Let $\pi = (\pi_k)_{k\in\mathbb{N}_0}$ be a sequence of measurable functions $\pi_k : H_k \times A \to [0,1]$ so that $\sum_{a\in D(x_k)} \pi_k(h_k,a) = 1$ and $\pi_k(h_k,a) = 0$ for all $a \notin D(x_k)$ and for all $k \in \mathbb{N}_0$ where $h_k = (x_0,a_0,x_1,\dots,a_{k-1},x_k) \in H_k$. Then π is called a *history-dependent randomized policy.* Let Π be the set of all history-dependent randomized policies. A policy $\pi \in \Pi$ is a *Markovian randomized policy* if $\pi_k(h_k,a) = \pi_k((x_0',a_0',x_1',\dots,a_{k-1}',x_k),a)$ for all $h_k = (x_0,a_0,x_1,\dots,a_{k-1},x_k) \in H_k$, and $x_0',\dots,x_{k-1}' \in S$, $a \in D(x_k)$, $a_l' \in D(x_l)$, $l = 0,\dots,k-1$, and for all $k \in \mathbb{N}_0$. The set of all Markovian randomized policies is denoted by Π_m. Moreover, let $\mu : D \to [0,1]$ such that $\sum_{a\in D(s)} \mu(s,a) = 1$ for all $s \in S$. Then the sequence $\mu^\infty := (\mu,\mu,\dots)$ is called a *stationary policy.* The set of all stationary policies be denoted by Π_s. We also write $\mu(a|s) := \mu(s,a)$ and identify μ with μ^∞ and also write $\mu \in \Pi_s$. A policy $\mu^\infty \in \Pi_s$ is a *deterministic stationary policy* if $\mu(s,a) \in \{0,1\}$ for all $(s,a) \in D$. If $\mu(s,a) = 1$ for some $(s,a) \in D$, then we also write $\mu(s) = a$. The set of all deterministic stationary policies is denoted by Π_s^d.

Let the sample space be $\Omega := \bigtimes_{i=0}^{\infty}(S \times A)$ and $\mathscr{A} := \bigtimes_{i=0}^{\infty}(\mathscr{P}(S) \times \mathscr{P}(A))$ be a σ-algebra on Ω. For an element $(x_0, a_0, x_1, a_1, \ldots) \in \Omega$, we define the projections $X_k(\omega) = x_1$ and $X_k(\omega) = a_k$. The processes $(X_k)_{k\in\mathbb{N}_0}$ and $(A_k)_{k\in\mathbb{N}_0}$ are the state process and the action process respectively. In this section, we only treat the case in which the initial state is known with probability one. Therefore, let P_0 be an initial distribution on $\mathscr{P}(S)$ with $P_0(\{s\}) = 1$ for a fixed initial state $s \in S$. For $\pi \in \Pi$, we can define a probability measure P^π on $(\Omega, \mathscr{A})$ by

$$
\begin{aligned}
&P^\pi(X_0 = x_0) = P_0(\{x_0\}),\\
&P^\pi\left(A_k = a \,|\, X_0 = x_0, A_0 = a_0, \ldots, X_k = x_k\right) = \pi_k(h_k, a),\\
&P^\pi\left(X_k = s' \,\middle|\, X_0 = x_0, A_0 = a_0, X_1 = x_1, \ldots, X_k = x_k, A_k = a_k\right) = p^a_{x_k s'},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x_0, \ldots, x_k, s' \in S,\ a_0, \ldots, a_k, a \in A,\ k \in \mathbb{N}_0.
\end{aligned}
$$

The probability measure P^π again exists and is unique by a theorem of Ionescu-Tulcea (cf. (Bertsekas and Shreve, 1978), Proposition 7.28).

In this section, we define the average cost by

$$C := \limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} c(X_k, A_k),$$

where $(X_k)_{k\in\mathbb{N}_0}$ is the state process and $(A_k)_{k\in\mathbb{N}_0}$ is the action process. Under every policy $\pi \in \Pi_s$, the average cost C is a random variable on $(\Omega, \mathscr{A})$. By this definition, the decision maker takes a conservative point of view by minimizing the pessimistic outcome of the average cost by considering the lim sup. The problem we want to solve for a given $\tau \in (0,1)$ and a given initial state $x_0 \in S$ is:

$$\text{Minimize } \mathrm{AV@R}^{\mu}_{\tau}(C \,|\, X_0 = x_0) \text{ over all } \mu \in \Pi_s. \tag{8.4}$$

In this chapter, optimality is defined in the following way unless otherwise stated.

Definition 8.4.3. Let $\tau \in (0,1)$ and $x_0 \in S$. A policy $\mu^* \in \Pi_s$ is called *optimal with respect to the average value-at-risk criterion at confidence level τ for the average cost (or* $\mathrm{AV@R}_\tau$*-optimal)* if $\mathrm{AV@R}^{\mu^*}_{\tau}(C \,|\, X_0 = x_0) = \inf_{\mu\in\Pi_s} \mathrm{AV@R}^{\mu}_{\tau}(C \,|\, X_0 = x_0)$.

Here, the property of optimality of a policy is associated with the initial state x_0. Here, we try to optimize the average value-at-risk within the class of all stationary policies.

8.5. Continuous- and Discrete-Time Markov Decision Processes

As in the discounted case, one can apply a uniformization technique in order to transform the continuous-time model into a discrete-time model such that the expected long-run average costs coincide in both models for every deterministic stationary policy. This technique is also provided by (Serfozo, 1979). But it is a bit different from the discounted case. Again, let $\tilde{\lambda}(s,a) := \tilde{\lambda} := \max_{(s',a')\in D} \lambda(s',a')$, $(s,a) \in D$, be the transition rate for all state-action pairs. The transition probabilities and the costs in the uniformized model are defined by

$$
\begin{aligned}
\tilde{p}^a_{ss'} &:= \begin{cases} 1 - \frac{\lambda(s,a)\,(1-p^a_{ss})}{\tilde{\lambda}}, & s' = s \\ \frac{\lambda(s,a)\,p^a_{ss'}}{\tilde{\lambda}}, & s' \neq s \end{cases},\\
\tilde{c}(s,a) &:= \frac{\lambda(s,a)}{\tilde{\lambda}}\, c(s,a),
\end{aligned}
$$

$s, s' \in S$ and $a \in D(s)$. (Beutler and Ross, 1987) show that this uniformization procedure does not result in the same expected average cost for the continuous-time and the uniformized model for general stationary policies. They refine the uniformization step such that the expected long-run average cost of the continuous-time and the uniformized model coincide. Nevertheless, a stationary optimal policy for the uniformized model is also optimal for the original model. Furthermore, techniques for solving discrete-time MDPs can be applied to solve the uniformized model. But there is no evidence that the uniformization procedure can be applied to the average value-at-risk criterion without any further consideration. However, we assume that this procedure can be applied without great loss when the data of the surveillance task are transformed to the discrete-time model.

8.6. Unichain and Weakly Communicating MDPs

At first, we show that a stationary expectation optimal policy is AV@R$_\tau$-optimal if the MDP is unichain. In the following, we obtain the same result for weakly communicating MDPs. In these cases, we can verify that there are deterministic stationary policies which are optimal in the class of all history-dependent randomized policies.

Definition 8.6.1. An MDP is called *unichain* if for every $\mu \in \Pi_s$ the corresponding Markov chain $(X_k)_{k\in\mathbb{N}_0}$ consists of exactly one class of recurrent states and additionally of a class of transient states, which could be empty.

Theorem 8.6.2. *Let Γ be a unichain MDP. Further, let $\mu^* \in \Pi_s$ be a stationary policy which is optimal with respect to the expected average cost criterion. Then μ^* is optimal with respect to the* AV@R$_\tau$*-criterion within the class of all history-dependent randomized policies for all $\tau \in (0,1)$ and for all $x_0 \in S$. Moreover, there exists a deterministic stationary policy $\mu^* \in \Pi_s^d$ which is* AV@R$_\tau$*-optimal within the class of all history-dependent randomized policies for all $\tau \in (0,1)$ and for all $x_0 \in S$.*

Proof. At first, there exists a deterministic stationary policy $\mu^* \in \Pi_s^d$ which is optimal with respect to the expected average cost criterion for all $x_0 \in S$ by (Puterman, 2005), Theorem 8.4.5. Now, let $\mu^* \in \Pi_s$ be an arbitrary stationary expected average cost optimal policy. Since Γ is unichain and $\mu^* \in \Pi_s$, the resulting Markov chain under μ^* consists of exactly one closed irreducible recurrent class and a possibly empty set of transient states. Hence, there is a unique stationary distribution π_{μ^*} on S. So, for all initial states $x_0 \in S$ one has

$$C = \limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} c(X_k, \mu^*(X_k)) = \sum_{s\in S} \pi_{\mu^*}(s)\, c(s, \mu^*(s)) \quad P^{\mu^*}\text{-a. s.}$$

by an ergodic theorem for homogeneous Markov chains (cf. (Brémaud, 1999), Theorem 4.1 of chapter 3.4). Hence, C is constant P^{μ^*}-a. s. for every $x_0 \in S$. Thus, for every $\tau \in (0,1)$

$$\text{AV@R}_\tau^{\mu^*}(C\,|\,X_0 = x_0) = E^{\mu^*}[C\,|X_0 = x_0] \le E^{\pi}[C\,|X_0 = x_0] \le \text{AV@R}_\tau^{\mu}(C\,|\,X_0 = x_0) \quad (\pi \in \Pi).$$

The last inequality follows from the fact that for a random variable X with $E|X| < \infty$, we have $EX \le \text{AV@R}_\tau(X)$ for every $\tau \in (0,1)$. □

By Theorem 8.6.2, considering unichain MDPs with respect to the average value-at-risk criterion for the average cost only requires theory on unichain MDPs with respect to the expected average cost criterion. A survey of such problems and its solution methods is given in (Puterman, 2005), sections 8.4–8.8. Since there always exists a deterministic stationary optimal policy with respect to the expected average cost criterion, there exists a deterministic stationary optimal policy for the AV@R$_\tau$-criterion. For its computation, the common Bellman equation for the expected average cost criterion can be used, which is in the case of a unichain MDP

$$g(s) + h(s) = \min_{a\in D(s)} \left\{ c(s,a) + \sum_{s'\in S} p_{ss'}^a\, h(s') \right\}, \quad s \in S, \tag{8.5}$$

and a minimizer $\mu^*(s)$ of (8.5) is optimal in state $s \in S$ by Theorem 8.4.4 of (Puterman, 2005). In practice, methods such as Howard's policy improvement algorithm and a linear programming approach are available for exact computation, and also approximate methods such as value iteration can be used to solve the problem.

The results of this section can also be applied to so-called weakly communicating MDPs, which we define as in (Puterman, 2005).

Definition 8.6.3. An MDP is *weakly communicating* if there is a closed set of states, with each state in that set accessible from every other state in that set under some deterministic stationary policy, plus a possibly empty set of states which is transient under every policy.

Theorem 8.6.4. *Let Γ be a weakly communicating MDP. Further, let $\mu^* \in \Pi_s$ be a stationary policy which is optimal with respect to the expected average cost criterion. Then μ^* is optimal with respect to the* AV@R$_\tau$*-criterion within the class of all history-dependent randomized policies for all $\tau \in (0,1)$ and for all $x_0 \in S$. Moreover, there exists a deterministic stationary policy $\mu^* \in \Pi_s^d$ which is* AV@R$_\tau$*-optimal within the class of all history-dependent randomized policies for all $\tau \in (0,1)$ and for all $x_0 \in S$.*

Proof. To obtain the results, first note that there is a deterministic stationary optimal policy with respect to the expected average cost criterion for all $x_0 \in S$ by (Puterman, 2005), Theorem 9.1.8, for a general finite-state finite-action MDP. Now, let $\mu^* \in \Pi_s$ be a stationary optimal policy with respect to the expected average cost criterion. Assuming a weakly communicating MDP, we get by Theorem 8.3.2 that $E^{\mu^*}[C\,|\,X_0 = x_0] = c$ for all $x_0 \in S$ for some $c \in \mathbb{R}$ independent of x_0. Each closed irreducible recurrent class $\mathscr{R}$ under μ^* must have a stationary distribution $\pi_{\mathscr{R}}$ restricted to the states of $\mathscr{R}$ such that $\sum_{s\in\mathscr{R}} \pi_{\mathscr{R}}(s)\, c(s,\mu^*(s)) = c$. Otherwise, there is a closed irreducible recurrent class $\mathscr{R}_1$ so that $\sum_{s\in\mathscr{R}_1} \pi_{\mathscr{R}_1}(s)\, c(s,\mu^*(s)) < c$. Then by a slight generalization of (Puterman, 2005), Theorem 8.3.2a, there is some $\hat{\mu} \in \Pi_s$ so that $\mathscr{R}_1$ is the only closed irreducible recurrent class and $\hat{\mu}(s,a) = \mu^*(s,a)$ for all $s \in \mathscr{R}_1$ and for all $a \in D(s)$, leading to $E^{\hat{\mu}}[C\,|\,X_0 = x_0] < c$, which is a contradiction to the optimality of μ^*. Since S is finite, we have $P^{\mu^*}(C = c\,|\,X_0 = x_0) = 1$ for all $x_0 \in S$. Therefore, as in the unichain case, we conclude

$$c = \mathrm{AV@R}_\tau^{\mu^*}(C\,|\,X_0 = x_0) = E^{\mu^*}[C\,|\,X_0 = x_0] \le E^{\pi}[C\,|\,X_0 = x_0] \le \mathrm{AV@R}_\tau^{\pi}(C\,|\,X_0 = x_0) \quad (\pi \in \Pi).$$

Hence, μ^* is optimal with respect to the $\mathrm{AV@R}_\tau$-criterion for all $\tau \in (0,1)$ and for all $x_0 \in S$. □

In weakly communicating MDPs, it is more cumbersome to obtain a deterministic stationary optimal policy with respect to the expected average cost criterion since theory for multichain MDPs has to be used in this case. Nevertheless, it is possible to find such policies by solving the "nested" optimality equations (cf. (Puterman, 2005))

$$\begin{aligned} &\min_{a\in D(s)} \left\{ \sum_{s'\in S} p_{ss'}^a\, g(s') - g(s) \right\} = 0, \\ &\min_{a\in B(s)} \left\{ c(s,a) - g(s) + \sum_{s'\in S} p_{ss'}^a\, h(s') - h(s) \right\} = 0, \quad s \in S, \end{aligned} \tag{8.6}$$

for $g, h : S \to \mathbb{R}$, where $B(s) \coloneqq \left\{ a \in D(s) : \sum_{s'\in S} p_{ss'}^a\, g(s') - g(s) = 0 \right\}$. By Theorems 9.1.7 and 9.1.8, a deterministic stationary optimal policy μ^* can be obtained in the following way: compute $g, h : S \to \mathbb{R}$ which satisfy (8.6) and choose $\mu^* \in \Pi_s^d$ such that $\sum_{s'\in S} p_{ss'}^{\mu^*(s)}\, g(s') = g(s)$ for all $s \in S$ and $\mu^*(s) \in \arg\min_{a\in B(s)} \left\{ c(s,\mu^*(s)) + \sum_{s'\in S} p_{ss'}^{\mu^*(s)}\, h(s') \right\}$ for all $s \in S$. In practice, algorithms like a linear programming approach and a policy improvement algorithm or approximate methods like value iteration are available, which differ from the respective variants for unichain MDPs.

8.7. Multichain MDPs

Throughout this section, we consider general finite-state finite-action MDPs. The goal is to find a method for determining $\mathrm{AV@R}_\tau^*(C\,|\,X_0 = x_0) \coloneqq \inf_{\mu\in\Pi_s} \mathrm{AV@R}_\tau^{\mu}(C\,|\,X_0 = x_0)$ for some given $x_0 \in S$ and some stationary policy $\mu^* \in \Pi_s$ for which $\mathrm{AV@R}_\tau^{\mu^*}(C\,|\,X_0 = x_0) = \mathrm{AV@R}_\tau^*(C\,|\,X_0 = x_0)$ holds. Before we come to that, we make a definition concerning the structure of a given MDP, which was first given in (Ross and Varadarajan, 1991). They treat the case of minimizing the expected average cost along a sample path constraint. In their task, together with rewards for choosing an action in a state costs arise, too. In their article, the constraint is that the average cost should stay below a given real value with probability one. Their approach is to decompose the state space into sets the elements of which have helpful properties. This approach is also useful in other non-standard criteria for the average cost, like for example in (Baykal-Gürsoy and Ross, 1992; Filar et al., 1995).

Definition 8.7.1. A class $\mathscr{R} \subset S$ is *strongly communicating*

1. if there is a $\mu \in \Pi_s$ such that $\mathscr{R}$ is closed irreducible and recurrent under μ and
2. if there is no $\mathscr{R}' \supsetneq \mathscr{R}$ such that $\mathscr{R}'$ satisfies 1.

In other words, a class $\mathscr{R}$ is strongly communicating if it is closed irreducible recurrent under some stationary policy and if there is no $\mathscr{R}' \supsetneq \mathscr{R}$ which is closed irreducible recurrent under some other stationary policy. Such a stationary policy may not be deterministic in general, as the next example shows.

Example 8.7.2. Let $S = \{1,2,3\}$, $A = \{1,2\}$ with $D(1) = D(3) = \{1\}$, $D(2) = \{1,2\}$ and transition probabilities

$$p_{12}^1 = 1, \quad p_{21}^1 = 1, \quad p_{23}^2 = 1, \quad p_{32}^1 = 1,$$

where all other transition probabilities are zero. Figure 8.1 illustrates this model. There is exactly one strongly communicating class, which equals S. Under every strictly randomized stationary policy, S is recurrent. Whereas for the two deterministic stationary policies, which either choose action 1 or action 2 in state 2 with probability one, there is exactly one transient state.

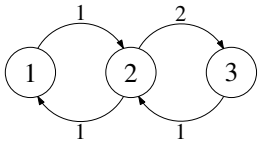


Figure 8.1.: MDP model of Example 8.7.2.

In Proposition 1, (Ross and Varadarajan, 1991) show that the classes $\mathscr{T}$ of states which are transient under all policies, and the strongly communicating classes $\mathscr{R}_1, \ldots, \mathscr{R}_R$ for some $R \in \mathbb{N}$, form a partition of S. Moreover, they give an algorithm for obtaining this partition. Another important property of a strongly communicating class is also provided in (Ross and Varadarajan, 1991), Lemma 2:

Lemma 8.7.3. *Let $\mathscr{R}$ be a strongly communicating class. Then for every $\pi \in \Pi$ and for every $x_0 \in S$*

$$P^{\pi}\left(\{X_n \in \mathscr{R} \text{ for infinitely many } n \in \mathbb{N}_0\} \cap \{X_n \notin \mathscr{R} \text{ for infinitely many } n \in \mathbb{N}_0\} \mid X_0 = x_0\right) = 0.$$

Proof. See (Ross and Varadarajan, 1991). □

Note that in Lemma 8.7.3 the policy $\pi \in \Pi$ need not be stationary. Further below, we fall back on the next definition. Let $s \in \mathscr{R}$, for some strongly communicating class $\mathscr{R} \subset S$. Then define

$$D_s := \left\{ a \in D(s) : \sum_{s' \in \mathscr{R}} p_{ss'}^{a} = 1 \right\}$$

as the set of actions the decision maker can choose in s after which the state transitions to a state also lying in $\mathscr{R}$ with probability one. If the MDP has a certain structure containing the next definition, then we are able to obtain optimal stationary policies quite easily, as we shall see.

Definition 8.7.4. A strongly communicating class $\mathscr{R}$ is called a *sink* if $D_s = D(s)$ for all $s \in \mathscr{R}$.

By this definition, a sink is a strongly communicating class which cannot be left under any policy, once it has been entered. To get an idea of sinks, we prove the following proposition:

Proposition 8.7.5. *Let Γ be an MDP. Then there exists a sink.*

Proof. Let $\mathscr{T}$ be the set of states which are transient under every policy. Further, let $\mathscr{R}_1, \ldots, \mathscr{R}_R$ be the strongly communicating classes. Assume that there is no sink. Hence, for all $r \in \{1, \ldots, R\}$, there is a state $s_r \in \mathscr{R}_r$, and there is an action $a_r \in D(s_r)$ such that $p_{s_r s'}^{a_r} > 0$ for some $s' \notin \mathscr{R}_r$. For every $r \in \{1, \ldots, R\}$, let μ_r be a stationary policy such that $\mathscr{R}_r$ is closed irreducible and recurrent under μ_r in view of the definition of strongly communicating classes. Then define a stationary policy $\mu \in \Pi_s$ by

$$\mu(s,a) := \begin{cases} \mu_1(s,a), & \text{if } s \in \mathscr{T} \\ \mu_r(s,a), & \text{if } s \in \mathscr{R}_r,\ s \neq s_r,\ r = 1, \ldots, R \\ \frac{1}{2}, & \text{if } s \in \mathscr{R}_r,\ s = s_r,\ a = a_r,\ r = 1, \ldots, R \\ \frac{1}{2}\mu_r(s,a), & \text{if } s \in \mathscr{R}_r,\ s = s_r,\ a \neq a_r,\ r = 1, \ldots, R \end{cases}$$

for $(s,a) \in D$. Under μ, all states in the strongly communicating class $\mathscr{R}_r$ are communicating. Furthermore, none of the strongly communicating classes is closed under μ since in all strongly communicating classes, there is one state which can be left with positive probability towards a different strongly communicating class or an element of $\mathscr{T}$. But since S is finite, there has to be a closed irreducible recurrent class. This class has to be of the form $\mathscr{R}_{r_1} \cup \ldots \cup \mathscr{R}_{r_{\tilde{R}}} =: \mathscr{R}$, $r_1, \ldots, r_{\tilde{R}} \in \{1, \ldots, R\}$, for some $\tilde{R} \geq 2$, w. l. o. g., since all states in $\mathscr{R}_r$ are communicating. Hence, $\mathscr{R}$ is closed irreducible and recurrent under μ with $\mathscr{R} \supsetneq \mathscr{R}_r$ for all $r \in \{r_1, \ldots, r_{\tilde{R}}\}$, which contradicts the assumption that the $\mathscr{R}_r$ are strongly communicating classes, which concludes the proof. □

The idea of finding an AV@R$_\tau$-optimal policy is similar to the one in (Ross and Varadarajan, 1991). At first, we solve the MDP just for the strongly communicating classes. Then we find an overall-policy which adequately finds a way through the transient states and the strongly communicating classes so that it is absorbed by the strongly communicating classes such that the distribution of the average cost minimizes the average value-at-risk of the average cost.

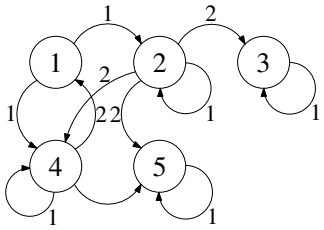


Figure 8.2.: Example of an MDP with strongly communicating classes consisting of one state only.

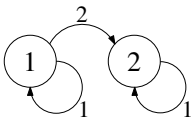


Figure 8.3.: MDP model of Example 8.7.6.

8.7.1. MDPs with Communicating Classes Consisting of Exactly One State

In this section, we restrict ourselves to a simpler kind of MDPs. In the following, we assume that the strongly communicating classes consist of exactly one state. Then a strongly communicating class $\mathscr{R}_r = \{r\}$, $r = 1, \ldots, R$, consists of a state which has some action $a_r \in D(r)$ leading back to r with probability one. Further, we assume that there is exactly one such action. But it is not excluded that there might be other actions that lead to other states of the MDP which might be either transient under every stationary policy or might be a strongly communicating class itself. We define

$$c(s) \coloneqq \begin{cases} 0, & \text{if } s \in \mathscr{T} \\ c(s,a_s), & \text{if } s \in \{1,\ldots,R\} \end{cases},$$

where $\mathscr{T}$ is the set of states which are transient under every stationary policy. Note that by Proposition 8.7.5 there exists at least one sink, which means that there is at least one state which cannot be left under any policy. Figure 8.2 gives an example of an MDP which meets the aforementioned assumptions without any parameters specified. The numbers on the arrows denote actions and arrows indicate strictly positive transition probabilities. State 1 is transient under every policy, whereas the states 2, 3, 4 and 5 are strongly communicating of which the states 3 and 5 are sinks.

For $\tau \in (0,1)$ and $x_0 \in S$, we want to solve problem (8.4). The set Π_s is compact since it can be identified with a closed subset of $[0,1]^n$ for $n = \sum_{s\in S} |D(s)|$ where components belonging to state s must add up to 1. But the functional $\mathrm{AV@R}_\tau^\mu(C \,|\, X_0 = x_0)$ is not continuous in μ with respect to any vector norm on $\mathbb{R}^n$, which is demonstrated by the following small example.

Example 8.7.6. Let $S = \{1,2\}$, $A = \{1,2\}$, $D(1) = \{1,2\}$, $D(2) = \{1\}$ with $p_{11}^1 = p_{12}^2 = p_{22}^1 = 1$ and $c(1) = 0$, $c(2) = 1$. A sketch of this model can be seen in Figure 8.3. Define dependent on $t \in [0,1]$ the stationary policy μ_t by $\mu_t(1,1) = t$, $\mu_t(1,2) = 1-t$ and $\mu_t(2,1) = 1$. Let $\tau \in (0,1)$ be arbitrary. Then

$$P^{\mu_t}\left((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } 1 \,\middle|\, X_0 = 1\right) = \begin{cases} 0, & \text{if } t \in [0,1) \\ 1, & \text{if } t = 1 \end{cases},$$

hence

$$\mathrm{AV@R}_\tau^{\mu_t}(C \,|\, X_0 = 1) = \begin{cases} 1, & \text{if } t \in [0,1) \\ 0, & \text{if } t = 1 \end{cases},$$

which yields the discontinuity of $\mathrm{AV@R}_\tau^{\mu_t}(C \,|\, X_0 = 1)$ in t and so in μ with respect to any vector norm on Π_s.

So, it is not clear from the beginning whether there exists an optimal policy amongst the stationary policies. Looking at the absorption probabilities and with Theorem 6.2.3 in mind, we consider the following optimization problem (8.7) for a given initial state $x_0 \in S$, which is examined in the following:

$$\text{Minimize } x + \frac{1}{1-\tau} \sum_{s'=1}^{R} \left(c(s') - x\right)^+ \alpha(x_0, s') \tag{8.7a}$$

under the constraints

$$x \in \mathbb{R}, \tag{8.7b}$$
$$0 \leq \mu(s,a) \leq 1 \quad ((s,a) \in D), \tag{8.7c}$$
$$0 \leq \alpha(s,s') \leq 1 \quad (s, s' \in S), \tag{8.7d}$$
$$\sum_{a \in D(s)} \mu(s,a) = 1 \quad (s \in S), \tag{8.7e}$$
$$\sum_{s' \in S} \alpha(s,s') = 1 \quad (s \in S), \tag{8.7f}$$
$$\alpha(s,s') = \sum_{a \in D(s)} \mu(s,a) \sum_{s'' \in S} p^a_{ss''}\, \alpha(s'',s') \quad (s, s' \in S), \tag{8.7g}$$
$$\alpha(s,s) = \lfloor \mu(s,a_s) \rfloor \quad (s \in \{1,\dots,R\}), \tag{8.7h}$$
$$\alpha(s,s') = 0 \quad (s \in S,\, s' \in \mathscr{T}). \tag{8.7i}$$

In this problem, $\lfloor x \rfloor$ denotes the greatest integer less than or equal to $x \in \mathbb{R}$, i. e., $\lfloor \cdot \rfloor$ is the floor function. The variables of problem (8.7) are x, $\alpha(s,s')$ and $\mu(s,a)$ for $s, s' \in S$ and $a \in D(s)$. If $\alpha(x_0,\cdot)$ is a fixed distribution on S, then minimization of the objective function in (8.7a) over $x \in \mathbb{R}$ yields the average value-at-risk of a random variable X with distribution $P(X = c(s)) = \sum_{s' \in \{1,\dots,R\}: c(s') = c(s)} \alpha(x_0,s')$ by the average value-at-risk representation theorem 6.2.3. Constraints (8.7c)–(8.7f) ensure that $\alpha(s,\cdot)$ and $\mu(s,\cdot)$ are probability measures on S and $D(s)$ for every $s \in S$ respectively. Constraint (8.7g) comes from the first-step analysis for homogeneous Markov chains. Constraints (8.7h) and (8.7i) give absorption probabilities for the strongly communicating states and the transient states respectively. A strongly communicating class which is not a sink could either be transient or recurrent under appropriate policies. So starting in it, either the state process will never return to it after a finite number of time-steps or it will be absorbed by it by Lemma 8.7.3 under any policy.

Remark 8.7.7. Observing problem (8.7), one recognizes that it is neither a convex nor a differentiable problem, which makes the computation of a solution hard. It is not a convex problem because of constraint (8.7g) and it is not a differentiable problem because of the objective function and the constraint (8.7h), which is not even continuous. However, problem (8.7) turns out to be useful in order to derive some theoretic results.

Proposition 8.7.8. *Let $(\mu(s,a))_{(s,a) \in D}$ be fixed so that $0 \leq \mu(s,a) \leq 1$ for all $(s,a) \in D$ and $\sum_{a \in D(s)} \mu(s,a) = 1$ for all $s \in S$. Then it holds for $x_0 \in S$*

$$\mathrm{AV@R}^{\mu}_{\tau}(C \,|\, X_0 = x_0) = \min(8.8),$$

where (8.8) *is the following optimization problem:*

$$\textit{Minimize } x + \frac{1}{1-\tau} \sum_{s'=1}^{R} \big(c(s') - x\big)^+ \alpha(x_0,s') \tag{8.8a}$$

under the constraints

$$x \in \mathbb{R}, \tag{8.8b}$$
$$0 \leq \alpha(s,s') \leq 1 \quad (s, s' \in S), \tag{8.8c}$$
$$\sum_{s' \in S} \alpha(s,s') = 1 \quad (s \in S), \tag{8.8d}$$
$$\alpha(s,s') = \sum_{a \in D(s)} \mu(s,a) \sum_{s'' \in S} p^a_{ss''}\, \alpha(s'',s') \quad (s, s' \in S), \tag{8.8e}$$
$$\alpha(s,s) = \lfloor \mu(s,a_s) \rfloor \quad (s \in \{1,\dots,R\}), \tag{8.8f}$$
$$\alpha(s,s') = 0 \quad (s \in S,\, s' \in \mathscr{T}). \tag{8.8g}$$

Proof. Since μ is a stationary policy, it defines a homogeneous Markov chain $(X_k,A_k)_{k \in \mathbb{N}_0}$ on the augmented state space $S \times A$ with initial distribution $P^{\mu}((X_0,A_0) = (x_0,a_0)) = \mu(x_0,a_0)$, $a_0 \in D(x_0)$. Together with the costs $c(s,a)$, $(s,a) \in D$, we have a Markov reward process defined by μ. So, we can use the result from Theorem 8.3.5, that the distribution of the average cost is given by the absorption probabilities in recurrent classes of $S \times A$ under μ. At first, note that for $s \in \{1,\dots,R\}$ we have

$$(s,a_s) \text{ is recurrent under } \mu \quad \Leftrightarrow \quad \mu(s,a_s) = 1 \quad \Leftrightarrow \quad s \text{ is recurrent under } \mu, \tag{8.9}$$

since the strongly communicating classes consist of exactly one state. Next, we verify that every solution α^* to problem (8.8) satisfies

$$\alpha^*(s,s') = P^\mu\left((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } s' \,|\, X_0 = s\right), \quad s, s' \in S, \tag{8.10}$$

which is true if the boundary conditions of the absorption probabilities are correct since (8.8e) is the first-step analysis. Constraints (8.8c) and (8.8d) assure that $\alpha^*(s,\cdot)$ is a probability measure on S for every $s \in S$. From (8.8f), we have $\alpha^*(s,s) = 1 \Leftrightarrow \mu(s,a_s) = 1$ and $\alpha^*(s,s) = 0 \Leftrightarrow \mu(s,a_s) < 1$ for every $s \in \{1,\dots,R\}$. This gives the correct absorption probabilities under μ since under μ a state $s \in S$ is recurrent if and only if $\mu(s,a_s) = 1$ and s is transient if and only if $\mu(s,a_s) < 1$. Furthermore, for a transient state $s' \in \mathscr{T}$ under μ, we have $\alpha^*(s,s') = 0$, which is given by constraint (8.8g). Hence, the variables $\alpha^*(s,s')$ are uniquely defined for every $s, s' \in S$. For an arbitrary $x \in \mathbb{R}$, we compute

$$\begin{aligned}
E^\mu\left[(C-x)^+ \,\middle|\, X_0 = x_0\right] &= E^\mu\left[\left(\limsup_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} c(X_k,A_k) - x\right)^+ \,\middle|\, X_0 = x_0\right] \\
&= \sum_{a_0\in D(x_0)} P^\mu\left(A_0 = a_0 \,|\, X_0 = x_0\right) E^\mu\left[\left(\limsup_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} c(X_k,A_k) - x\right)^+ \,\middle|\, (X_0,A_0) = (x_0,a_0)\right] \\
&\overset{\text{Theorem 8.3.5}}{=} \sum_{a_0\in D(x_0)} P^\mu\left(A_0 = a_0 \,|\, X_0 = x_0\right) \\
&\qquad \cdot \sum_{r=1}^{R} (c(r,a_r) - x)^+ P^\mu\left((X_k,A_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } (r,a_r) \,\middle|\, (X_0,A_0) = (x_0,a_0)\right) \\
&= \sum_{r=1}^{R} (c(r) - x)^+ \sum_{a_0\in D(x_0)} P^\mu\left(A_0 = a_0 \,|\, X_0 = x_0\right) P^\mu\left((X_k,A_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } (r,a_r) \,\middle|\, (X_0,A_0) = (x_0,a_0)\right) \\
&= \sum_{r=1}^{R} (c(r) - x)^+ \sum_{a_0\in D(x_0)} P^\mu\left((X_k,A_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } (r,a_r), A_0 = a_0 \,\middle|\, X_0 = x_0\right) \\
&= \sum_{r=1}^{R} (c(r) - x)^+ P^\mu\left((X_k,A_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } (r,a_r) \,\middle|\, X_0 = x_0\right) \\
&\overset{(8.9)}{=} \sum_{r=1}^{R} (c(r) - x)^+ P^\mu\left((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } r \,\middle|\, X_0 = x_0\right) \overset{(8.10)}{=} \sum_{r=1}^{R} (c(r) - x)^+ \alpha^*(x_0,r).
\end{aligned} \tag{8.11}$$

Since in problem (8.8) all variables except x are fixed, as shown up to this point, we have

$$\begin{aligned}
\inf_{x\in\mathbb{R}}\left\{x + \frac{1}{1-\tau}\sum_{r=1}^{R} (c(r) - x)^+ \alpha^*(x_0,r)\right\} &\overset{(8.11)}{=} \inf_{x\in\mathbb{R}}\left\{x + \frac{1}{1-\tau} E^\mu\left[(C-x)^+ \,\middle|\, X_0 = x_0\right]\right\} \\
&\overset{\text{Theorem 6.2.3}}{=} \mathrm{AV@R}_\tau^\mu(C \,|\, X_0 = x_0),
\end{aligned}$$

which proves the assertion. □

With the last proposition in hand, we are able to prove the following statement. It links the optimization problem (8.7) with the original problem (8.4).

Theorem 8.7.9. *For $\tau \in (0,1)$ and $x_0 \in S$, it holds*

$$\inf_{\mu\in\Pi_s} \mathrm{AV@R}_\tau^\mu(C \,|\, X_0 = x_0) = \inf(8.7).$$

Furthermore, if there exists a solution μ^* to problem (8.7)*, then μ^* is $\mathrm{AV@R}_\tau$-optimal within the class of stationary policies.*

Proof. Since the constraints are the same, we have that problem (8.7) is equivalent to the problem:

$$\text{Minimize problem (8.8)} \tag{8.12a}$$

under the additional constraints

$$0 \le \mu(s,a) \le 1 \quad ((s,a) \in D), \tag{8.12b}$$

$$\sum_{a\in D(s)} \mu(s,a) = 1 \quad (s \in S). \tag{8.12c}$$

Since the constraints (8.12b) and (8.12c) determine the full space of stationary policies Π_s, we conclude the assertion from Proposition 8.7.8. □

Remark 8.7.10. 1. Theorem 8.7.9 guarantees that problem (8.7) is equivalent to problem (8.4). Unfortunately, we cannot conclude that there exists an AV@R$_\tau$-optimal stationary policy in general from problem (8.7).

2. The approach used to solve problem (8.4) could be transferred to other criteria for MDPs considering the average cost. This can be done by replacing the objective function by another one. For instance, to determine the minimal expected average cost, consider the objective function $\sum_{s'=1}^{R} c(s')\,\alpha(x_0,s')$. But since the problem is neither convex nor differentiable, this approach is not handier than the common linear programming approach for the expected average cost criterion.

If we got rid of the "bad," i. e., not differentiable, constraint (8.7h) of problem (8.7), then we could conclude that there indeed is an optimal stationary policy. For an MDP model consisting only of states which are transient under all stationary policies and of states which are sinks this constraint does not appear in the problem formulation. Therefore, we have the following corollary.

Corollary 8.7.11. *Let $\tau \in (0,1)$ and $x_0 \in S$. If the states $1,\dots,R$ are sinks, then there exists an optimal stationary policy.*

Proof. Since the states $1,\dots,R$ are sinks, the only possible action in state $s \in \{1,\dots,R\}$ is a_s. So, $\mu(s,a_s) = 1$ for all $s \in \{1,\dots,R\}$. Then constraint (8.7h) is equivalent to $\alpha(s,s) = 1$ for all $s \in \{1,\dots,R\}$. Since the constraints associated with the variables $\alpha(s,s')$ and $\mu(s,a)$ of problem (8.7) are differentiable, they define a closed set. Furthermore, these variables are bounded because of the constraints (8.7c) and (8.7d). Moreover, the objective function is continuous and tends to infinity as $x \to \pm\infty$ for all feasible $\alpha(s,s')$ and $\mu(s,a)$. Therefore, the assertion follows since a compact set of all variables can be constructed which contains the minimum point in a similar manner as in the proof of Proposition 7.5.16. □

Note that also with the assumptions of Corollary 8.7.11, the optimization problem (8.7) does not degenerate into a convex or differentiable problem because of the non-differentiable objective function and the non-convex constraint (8.7g).

Next, we prove the existence of an AV@R$_\tau$-optimal stationary policy if not all strongly communicating states are sinks. Its derivation is much more intricate than in the previous case. In its proof, it is crucial that S is supposed to be finite.

Theorem 8.7.12. *Let $\tau \in (0,1)$ and $x_0 \in S$. There exists an optimal stationary policy for problem* (8.4).

Proof. Let $\mathscr{T} \subset S$ be the set of states which are transient under every policy. Let $\mathscr{S} \subset S$ be the set of sinks. Furthermore, let $\mathscr{R} := S \setminus (\mathscr{T} \cup \mathscr{S})$ be the set of states which may be transient for some stationary policy and recurrent under another stationary policy. We have $s \in \mathscr{R}$ is recurrent if and only if $\mu(s,a_s) = 1$. The idea of the proof is to determine an optimal stationary policy under the constraint that a subset of states of $\mathscr{R}$ is recurrent and the complement is transient, and then to take the stationary policy according to such a subset which performs better than the optimal stationary policies arising from the other subsets. This goes well because S is finite. But we have to take care that every stationary policy is considered. Therefore, let $\tilde{\mathscr{S}} := \{s_1,\dots,s_{\tilde{R}}\} \subset \mathscr{R}$ be a possibly empty subset of states of $\mathscr{R}$ for some $\tilde{R} \in \mathbb{N}_0$ and modify the MDP Γ to an MDP $\Gamma_{\tilde{\mathscr{S}}}$ such that $D_{\tilde{\mathscr{S}}}(s) = \{a_s\}$ for all $s \in \tilde{\mathscr{S}}$, i. e., all states from $\tilde{\mathscr{S}}$ are sinks, and $D_{\tilde{\mathscr{S}}}(s) = D(s)$ for all $s \notin \tilde{\mathscr{S}}$. Furthermore, restrict the stationary policies so that $0 \le \mu(s,a_s) < 1$ for all $s \in \mathscr{R} \setminus \tilde{\mathscr{S}}$. In doing so, states lying in $\mathscr{R} \setminus \tilde{\mathscr{S}}$ are transient under every feasible stationary policy. (The case $\mu(s,a_s) = 1$, $s \in \mathscr{R} \setminus \tilde{\mathscr{S}}$, is considered within a problem for the choice of a different $\mathscr{R}$.) The related optimization problem is (8.13):

$$\begin{aligned}
&\text{Minimize } x + \frac{1}{1-\tau} \sum_{s' \in \mathscr{S} \cup \tilde{\mathscr{S}}} \big(c(s') - x\big)^+ \alpha(s,x_0) \\
&\text{under the constraints} \\
&x \in \mathbb{R}, \\
&\mu(s,a_s) = 1 \quad (s \in \mathscr{S} \cup \tilde{\mathscr{S}}), \\
&0 \le \mu(s,a) \le 1 \quad ((s,a) \in D), \\
&0 \le \mu(s,a_s) < 1 \quad (s \in \mathscr{R} \setminus \tilde{\mathscr{S}}), \\
&0 \le \alpha(s,s') \le 1 \quad (s, s' \in S), \\
&\sum_{a \in D_{\tilde{\mathscr{S}}}(s)} \mu(s,a) = 1 \quad (s \in S),
\end{aligned} \tag{8.13a}$$

$$\sum_{s' \in S} \alpha(s,s') = 1 \quad (s \in S),$$
$$\alpha(s,s') = \sum_{a \in D_{\tilde{\mathscr{S}}}(s)} \mu(s,a) \sum_{s'' \in S} p^a_{ss''}\, \alpha(s'',s') \quad (s,\, s' \in S),$$
$$\alpha(s,s) = 1 \quad (s \in \mathscr{S} \cup \tilde{\mathscr{S}}),$$
$$\alpha(s,s') = 0 \quad (s \in S,\, s' \in \mathscr{T} \cup (\mathscr{R} \setminus \tilde{\mathscr{S}})).$$

Let μ be feasible for problem (8.13). Then, as seen in the proof of Proposition 8.7.8, the $\alpha(s,s')$ are uniquely determined, and they are the absorption probabilities under policy μ. The policy $\hat{\mu}$ defined by

$$\hat{\mu}(s,a) := \begin{cases} \mu(s,a), & \text{if } s \in \mathscr{T} \cup \mathscr{S} \cup \tilde{\mathscr{S}},\, a \in D(s) \\ \frac{\mu(s,a)}{\sum_{a \in D_{\tilde{\mathscr{S}}} \setminus \{a_s\}} \mu(s,a)}, & \text{if } s \in \mathscr{R} \setminus \tilde{\mathscr{S}},\, a \neq a_s \\ 0, & \text{if } s \in \mathscr{R} \setminus \tilde{\mathscr{S}},\, a = a_s \end{cases}$$

has the same absorption probabilities as μ since a state $s \in \mathscr{R} \setminus \tilde{\mathscr{S}}$ is left with the same probability towards another state under $\hat{\mu}$ as s is left under μ. Note that $\hat{\mu}$ is also feasible for problem (8.13). Every μ which is feasible for problem (8.13) can be mapped onto such a feasible $\hat{\mu}$ which has the same objective function value for every $x \in \mathbb{R}$. So, constraint (8.13a) can be replaced by the constraint $\mu(s,a_s) = 0$ for all $s \in \mathscr{R} \setminus \tilde{\mathscr{S}}$ yielding an optimization problem the solutions of which are solutions to (8.13). The constraints of the equivalent problem describe a closed set of variables. Here again, the objective function tends to infinity as $x \to \pm\infty$ for every feasible α and μ in the equivalent optimization problem. Thus, there is a non-empty bounded subset of the set described by the constraints which contains points the objective function values of which are less than outside of this subset (cf. proof of Proposition 7.5.16). Since the objective function is continuous, there exists some $\tilde{\mu}$ which takes the minimum value on this subset, which then is a minimum of the set described by the constraints. Subsequently going through all subsets $\tilde{\mathscr{S}} \subset \mathscr{R} \setminus \mathscr{S}$ in the same way, every stationary policy is considered in one of the MDPs $\Gamma_{\tilde{\mathscr{S}}}$. Selecting one MDP $\Gamma_{\tilde{\mathscr{S}}}$ with minimal solution of the objective function and an associated optimal stationary policy yields the assertion since $\mathscr{R}$ is finite. □

Remark 8.7.13.

1. With the same method Theorem 8.7.12 is proven with, one can prove the existence of an optimal stationary policy for MDPs the strongly communicating classes of which consist of one state only if the objective function is continuous with respect to the absorption probabilities. For example, minimizing the value-at-risk at level $\tau \in (0,1)$ for such an MDP, leads to the objective function $\sum_{s'=1}^{R} \rho_\tau(c(s') - x)\, \alpha(x_0,s')$ where $\rho_\tau(x) = (\tau - 1)\, x$, $x < 0$, and $\rho_\tau(x) = \tau x$, $x \geq 0$. This relationship between a quantile and an expectation of the loss function ρ_τ is used for example in quantile regression as in (Koenker, 2005). The constraints remain the same as in problem (8.7). Therefore, the above method can be used, and hence, there exists an optimal stationary policy with respect to the value-at-risk criterion. Indeed, this problem was solved earlier in (Filar et al., 1995). Their result is also based on the decomposition approach. Furthermore, they consider intermediate MDPs with rewards 1 and 0 for the states of strongly communicating classes. Then maximizing the average reward of the intermediate MDPs, leads to maximizing the probability of ending up in those MDPs the rewards of which are set to 1. By this, the optimal value-at-risk can be found by solving the MDPs for the strongly communicating classes, ordering the strongly communicating classes by the amount of their average cost and then minimizing the probabilities of ending up in the strongly communicating with highest costs, until the probability of ending up in the strongly communicating classes with highest costs exceeds the given level $\tau \in (0,1)$. The approach we have chosen here is a little more intricate since we need more information about the distribution of the upper end of the distribution of the average cost.

2. Since the value-at-risk minimizes the term of the representation theorem of the average value-at-risk (Theorem 6.2.3), constraint (8.7b) can be replaced by the relaxed constraint $x \in \{c(s) : s \in S\}$ in order to determine the average value-at-risk of the average cost.

8.7.2. General MDPs

Having solved problem (8.4) under the assumption that the strongly communicating classes consist of exactly one state, the next step is to drop this assumption.

Now, we come to the main results of this section. They provide that there exist optimal stationary policies under certain assumptions for general finite-state finite-action MDPs. Otherwise, one could try to construct semi-stationary policies which are stationary in some states and history-dependent in others. Furthermore, we show that optimal stationary policies

Algorithm 8.1 Computation of an AV@R$_\tau$-optimal stationary or semi-stationary policy for multichain MDPs.

Require: Γ, $\tau \in (0,1)$, $x_0 \in S$
1: determine $\mathscr{R}_1, \dots, \mathscr{R}_R$, $\mu_1, \dots, \mu_R$ and $\mathscr{T}$
2: determine $\mathscr{S}$
3: **for** $r = 1$ to R **do**
4: determine $\mu_r^*(s,a)$ for $s \in \mathscr{R}_r$ and $c(r)$
5: **end for**
6: determine $\tilde{\Gamma} = (\tilde{S}, \tilde{A}, \tilde{D}, \tilde{P}, \tilde{\mathscr{C}})$
7: determine $\tilde{\mu}^*$ and $\tilde{\mathscr{S}}^*$
8: **if** $\tilde{\mu}^*(\tilde{s},\tilde{a}) \in \{0,1\}$ for all $\tilde{s} \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*$, $\tilde{a} \in \tilde{D}(\tilde{s})$ **then**
9: determine $\mu^* \in \Pi_s$
10: **else**
11: try to determine a semi-stationary policy which has the same average value-at-risk as $\tilde{\mu}^*$ in $\tilde{\Gamma}$
12: **end if**

can be constructed in the way we formerly proposed. We state Algorithm 8.1 and prove its correctness in Theorem 8.7.14. The details of the algorithm are explained in the following.

The input is a discrete-time MDP $\Gamma = (S, A, D, P, \mathscr{C})$ as introduced in the beginning of section 8.4, the confidence level $\tau \in (0,1)$ and some initial state $x_0 \in S$. Here, we assume that $D(s) \cap D(s') = \emptyset$ for all $s \neq s'$ so that every action is uniquely associated with one state. This is done to circumvent intricate notation w. l. o. g.. At first, in line 1, the strongly communicating classes $\mathscr{R}_r$, $r = 1, \dots, R$, of Γ shall be computed, together with associated stationary policies μ_r, $r = 1, \dots, R$, so that $\mathscr{R}_r$ is recurrent under μ_r. Furthermore, the set $\mathscr{T} \subset S$ of the states which are transient under all stationary policies shall be computed. This can be done by the partitioning algorithm presented in (Ross and Varadarajan, 1991). Next, the sinks of Γ are determined, which can be done by going through all states in a strongly communicating class and checking if every action ends up in the same strongly communicating class with probability one. Therefore, let $\mathscr{S} \subset \{1, \dots, R\}$ be the indices of the strongly communicating classes which are sinks. This is needed in line 7. In the for loop 3–5, stationary expected average cost optimal policies for the strongly communicating classes $\mathscr{R}_r$, $r = 1, \dots, R$, and the respective costs $c(r)$, which are the same for each state lying in $\mathscr{R}_r$, are computed. In line 6, a new MDP $\tilde{\Gamma}$ is generated such that the strongly communicating classes consist of one state only: the state space is $\tilde{S} = \{1, \dots, R\} \cup \mathscr{T}$. The restriction set is $\tilde{D}(\tilde{s}) := D(\tilde{s})$ for all $\tilde{s} \in \mathscr{T}$ and $\tilde{D}(r) := \{a \in D(s) : s \in \mathscr{R}_r,\ p_{ss'}^a > 0 \text{ for some } s' \notin \mathscr{R}_r\} \cup \{a_r\}$, $r = 1, \dots, R$, where $a_r \neq a_{r'}$, $r \neq r'$, and $a_r \notin A$ for all $r \in \{1, \dots, R\}$. All actions which could end up outside of the strongly communicating class $\mathscr{R}_r$ and an additional action a_r which forces the state process to remain in $r = 1, \dots, R$, are contained in $\tilde{D}(r)$. Then the canonical action space is $\tilde{A} := \bigcup_{\tilde{s} \in \tilde{S}} \tilde{D}(\tilde{s})$. The transition probabilities of $\tilde{\Gamma}$ are defined by

$$\tilde{p}_{\tilde{s}\tilde{s}'}^{\tilde{a}} := \begin{cases} p_{\tilde{s}\tilde{s}'}^{\tilde{a}}, & \text{if } \tilde{s}, \tilde{s}' \in \mathscr{T},\ \tilde{a} \in \tilde{D}(\tilde{s}) \\ \sum_{s' \in \mathscr{R}_{\tilde{s}'}} p_{\tilde{s}s'}^{\tilde{a}}, & \text{if } \tilde{s} \in \mathscr{T},\ \tilde{s}' \in \{1, \dots, R\},\ \tilde{a} \in \tilde{D}(\tilde{s}) \\ p_{ss'}^{\tilde{a}}, & \text{if } \tilde{s} \in \{1, \dots, R\},\ \tilde{s}' \in \mathscr{T},\ s \in \mathscr{R}_{\tilde{s}},\ \tilde{a} \in D(s) \\ \sum_{s' \in \mathscr{R}_{\tilde{s}'}} p_{ss'}^{\tilde{a}}, & \text{if } \tilde{s}, \tilde{s}' \in \{1, \dots, R\},\ s \in \mathscr{R}_{\tilde{s}},\ \tilde{a} \in D(s) \\ 1, & \text{if } \tilde{s} = \tilde{s}' \in \{1, \dots, R\},\ \tilde{a} = a_{\tilde{s}} \\ 0, & \text{else} \end{cases}.$$

The stochastic dynamics have not changed in $\tilde{\Gamma}$ for transient states. The states of a strongly communicating class $r \in \{1, \dots, R\}$ are assimilated in such a manner that the over-all dynamics are not modified by introducing an action a_r which forces the state process to remain in the strongly communicating class r representing a stationary policy so that the strongly communicating class is recurrent. The cost structure $\tilde{\mathscr{C}}$ shall be the following:

$$\tilde{c}(\tilde{s}) := \begin{cases} 0, & \text{if } \tilde{s} \in \mathscr{T} \\ c(\tilde{s}), & \text{if } \tilde{s} \in \{1, \dots, R\} \end{cases},$$

which is independent of the chosen action $\tilde{a} \in \tilde{D}(\tilde{s})$ for all $\tilde{s} \in \tilde{S}$. In line 7, an AV@R$_\tau$-optimal stationary policy $\tilde{\mu}^*$ of $\tilde{\Gamma}$ shall be computed and a subset $\tilde{\mathscr{S}}^* \subset \{1, \dots, R\}$ such that all states lying in $\tilde{\mathscr{S}}^*$ are sinks. Such a policy and the set $\tilde{\mathscr{S}}^*$ are provided by Theorem 8.7.12 since $\tilde{\Gamma}$ is of the form considered previously where all strongly communicating classes consist of exactly one state. The last step is putting together an AV@R$_\tau$-optimal stationary policy μ^* for the over-all

problem or a semi-stationary policy if possible. There are two cases: 1. $\tilde{\mu}^*$ is deterministic in the states $r \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*$ and 2. $\tilde{\mu}^*$ is randomized in one of these states. In the first case, we define a stationary policy $\mu^* \in \Pi_s$ by

$$\mu^*(s,a) \coloneqq \begin{cases} \tilde{\mu}^*(s,a), & \text{if } s \in \mathscr{T},\ a \in D(s) \\ \mu_r(s,a), & \text{if } s \in \mathscr{R}_r,\ \tilde{\mu}^*(r) \notin D(s),\ r \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*,\ a \in D(s) \\ 0, & \text{if } s \in \mathscr{R}_r,\ r \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*,\ a \neq \tilde{\mu}^*(r) \\ 1, & \text{if } s \in \mathscr{R}_r,\ r \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*,\ a = \tilde{\mu}^*(r) \\ \mu_r^*(s,a), & \text{if } s \in \mathscr{R}_r,\ r \in \tilde{\mathscr{S}}^*,\ a \in D(s) \end{cases}.$$

Note that μ^* need not be deterministic since $\tilde{\mu}^*$ need not be deterministic for some states $s \in \mathscr{T}$, and since μ_r need not be deterministic for some $r \in \tilde{\mathscr{S}}^*$ as Example 8.7.2 shows. In the second case, which is not excluded from consideration, one can try to compute a policy which leads to the same value of the average value-at-risk for $\tilde{\Gamma}$. In general, there need not exist a stationary policy with this property. In the states lying in $\mathscr{T}$ and $\mathscr{R}_r$, $r \in \tilde{\mathscr{S}}$, the definition of such a policy π^* is just the same as for μ^*. But once entered a strongly communicating class $\mathscr{R}_r$, $r \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*$, we must assure that the appropriate actions are chosen with the correct probabilities in order to leave the respective strongly communicating class with the probabilities given by $\tilde{\mu}^*$. This action choice is depending on the entry state and the time the MDP has already stayed in $\mathscr{R}_r$. This definition is quite cumbersome because, in general, in Γ actions with $\tilde{\mu}^*(r,a) > 0$ cannot be chosen from one state. If we had shown that for an MDP the strongly communicating classes of which consist of exactly one state there indeed is a deterministic optimal stationary policy, the second case would be obsolete.

Theorem 8.7.14. *Let $\Gamma = (S,A,D,P,\mathscr{C})$ be an MDP, $\tau \in (0,1)$ be the confidence level and $x_0 \in S$ be the initial state. Then Algorithm 8.1 computes an* $\mathrm{AV@R}_\tau$*-optimal stationary policy within the class of all stationary policies if $\tilde{\mu}^*(\tilde{s},\tilde{a}) \in \{0,1\}$ for all $\tilde{s} \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*$, $\tilde{a} \in \tilde{D}(\tilde{s})$.*

Proof. Let $\mathscr{R}_1,\dots,\mathscr{R}_R$ be the strongly communicating classes where, w. l. o. g., $\mathscr{S}$ is the set of the indices of the sinks. Further, let $\tilde{\mathscr{S}} \subset \{1,\dots,R\} \setminus \mathscr{S}$ be arbitrary. Now, we consider stationary policies the sinks of which are exactly the strongly communicating classes $\mathscr{R}_r$, $r \in \mathscr{S} \cup \tilde{\mathscr{S}}$. Let $\mu \in \Pi_s$ be such a policy. Moreover, consider the policy $\hat{\mu} \in \Pi_s$ defined by

$$\hat{\mu}(s,a) = \begin{cases} \mu(s,a), & \text{if } s \in \mathscr{T} \cup \bigcup_{r \in \{1,\dots,R\} \setminus (\mathscr{S} \cup \tilde{\mathscr{S}})} \mathscr{R}_r,\ a \in D(s) \\ \mu_r^*(s,a), & \text{if } s \in \mathscr{R}_r,\ r \in \mathscr{S} \cup \tilde{\mathscr{S}},\ a \in D(s) \end{cases}.$$

Then the probabilities of absorption by the strongly communicating classes $\mathscr{R}_r$, $r \in \mathscr{S} \cup \tilde{\mathscr{S}}$, under $\hat{\mu}$ are the same as under μ. Since every $\mathscr{R}_r$, $r \in \mathscr{S} \cup \tilde{\mathscr{S}}$, is a sink and all states in $\mathscr{R}_r$ are communicating, we have $P^\mu(C \geq c(r) \mid X_0 = s) = 1$ for every $s \in \mathscr{R}_r$ where $c(r) \coloneqq E^{\mu_r^*}[C \mid X_0 = s]$, which is the same for every $s \in \mathscr{R}_r$ by Theorems 8.3.2 and 9.1.8 of (Puterman, 2005). Because C takes only a finite number of values P^μ-a. s. by Theorem 8.3.5, we obtain from monotonicity of the average value-at-risk

$$\mathrm{AV@R}_\tau^\mu(C \mid X_0 = x_0) \geq \mathrm{AV@R}_\tau^{\hat{\mu}}(C \mid X_0 = x_0)$$

since we have $P^{\hat{\mu}}\left(C = c(r) \mid (X_k)_{k \in \mathbb{N}_0} \text{ is absorbed by } \mathscr{R}_r\right) = 1$ and $P^\mu\left(C \geq c(r) \mid (X_k)_{k \in \mathbb{N}_0} \text{ is absorbed by } \mathscr{R}_r\right) = 1$ for all $r \in \mathscr{S} \cup \tilde{\mathscr{S}}$. So, choosing actions according to the respective μ_r^* for all $s \in \bigcup_{r \in \mathscr{S} \cup \tilde{\mathscr{S}}} \mathscr{R}_r$, does not perform worse than choosing actions according to any μ modified in such a manner. Therefore, solving MDP $\tilde{\Gamma}$ leads to an $\mathrm{AV@R}_\tau$-optimal stationary policy μ^* as indicated in the detailed explanation of the algorithm. To this end, one has to remember that $s \in \mathscr{R}_r$ with $\tilde{\mu}^*(r) \in D(s)$ for some $r \in \{1,\dots,R\}$ (which means that $\mathscr{R}_r$ is transient under the proposed μ^* since $a_r \notin D(s)$ for any $s \in \mathscr{R}_r$ by assumption) is recurrent under μ_r, once $\mathscr{R}_r$ has been entered. Hence, it leaves the strongly communicating class $\mathscr{R}_r$ with probability $p_{ss'}^{\tilde{\mu}^*(s)}$ towards $s' \in \mathscr{T}$ and with probability $\sum_{s' \in \mathscr{R}_{r'}} p_{ss'}^{\tilde{\mu}^*(s)}$ towards the strongly communicating class $\mathscr{R}_{r'}$ for all $r' \in \{1,\dots,R\}$ which corresponds to the dynamics of $\tilde{\Gamma}$ under $\tilde{\mu}$. Together, μ^* is optimal under the given assumptions. □

Remark 8.7.15. Consider the second case of Algorithm 8.1, that is $\tilde{\mu}^*(\tilde{s},\tilde{a}) \notin \{0,1\}$ for some $\tilde{s} \in \{1,\dots,R\} \setminus \tilde{\mathscr{S}}^*$, $\tilde{a} \in \tilde{D}(\tilde{s})$. What makes this case much more intricate is the fact that the strongly communicating classes which are transient under $\tilde{\mu}^*$ in $\tilde{\Gamma}$ must be left with the same transition probabilities. Then we define another MDP $\hat{\Gamma} = (\hat{S},\hat{A},\hat{D},\hat{P},\hat{\mathscr{C}})$ with $\hat{S} = \mathscr{R}_r \cup \{\Delta_a : a \in D(s), s \in \mathscr{R}_r\}$ where the states Δ_a, are some absorbing states indicating that action a is chosen. The restrictions sets are $\hat{D}(\hat{s}) = D(\hat{s})$, $\hat{s} \in \mathscr{R}_r$, and $\hat{D}(\Delta_a) = \{\delta\}$, $\delta \notin A$, $a \in D(s)$, $s \in \mathscr{R}_r$, and $\hat{A} = \bigcup_{\hat{s} \in \hat{S}} \hat{D}(\hat{s})$. The transition probabilities are

$$\hat{p}_{\hat{s}\hat{s}'}^{\hat{a}} = \begin{cases} p_{\hat{s}\hat{s}'}^{\hat{a}}, & \text{if } \hat{s},\hat{s}' \in \mathscr{R}_r,\ \hat{a} \in D(\hat{s}),\ \tilde{\mu}^*(r,\hat{a}) = 0 \\ 1, & \text{if } \hat{s} \in \mathscr{R}_r,\ \hat{s}' = \Delta_{\hat{a}},\ \hat{a} \in D(\hat{s}),\ \tilde{\mu}^*(r,\hat{a}) > 0 \\ 1, & \text{if } \hat{s} = \hat{s}' \in \{\Delta_a : a \in D(s), s \in \mathscr{R}_r\},\ \hat{a} = \delta \\ 0, & \text{else} \end{cases}.$$

The costs are given by $\hat{c}(\hat{s},\hat{a}) = 0$, $\hat{s} \in \hat{S}$, $\hat{a} \in \hat{D}(\hat{s})$, since they do not play a role here. Assume there is a policy $\hat{\pi}_r \in \Pi$, not necessarily stationary, so that

$$P^{\hat{\pi}_r}\left((X_k)_{k\in\mathbb{N}_0} \text{ is absorbed by } \Delta_a \,\middle|\, X_0 = s\right) = \tilde{\mu}^*(r,a)$$

for all $s \in \mathscr{R}_r$ and $a \in \tilde{D}(r)$ for every $r \in \{1,\dots,R\}$.

Then we are able to construct a semi-stationary policy $\pi^* \in \Pi$ which has the same average value-of-risk as the one of $\tilde{\Gamma}$ under $\tilde{\mu}^*$. For all states in $s \in \mathscr{T}$, choose $\tilde{\mu}^*(s)$. If $\tilde{\mu}^*(r,a) \in \{0,1\}$ for all $a \in \tilde{D}(r)$ for some $r \in \{1,\dots,R\}$, then route the process through $\mathscr{R}_r$ (according to μ_r) until the state s' is reached such that $\tilde{\mu}^*(r) \in D(s')$. Then choose action $\tilde{\mu}^*(r)$ with probability one. If $0 < \tilde{\mu}^*(r,a) < 1$ for some $a \in \tilde{D}(r)$, $r \in \{1,\dots,R\}$, then act according to $\hat{\pi}_r$ until some action a with $\tilde{\mu}(r,a) > 0$ is chosen. If the subsequent state is again a state $s' \in \mathscr{R}_r$, then again act according to $\hat{\pi}_r$ starting at time step 0. By this, it is guaranteed that action a is chosen with probability $\tilde{\mu}(r,a)$ as long as the process remains in $\mathscr{R}_r$. Any other policy than π does not perform better than $\tilde{\mu}^*$ for the intermediate MDP $\tilde{\Gamma}$.

8.7.3. Example

We illustrate Algorithm 8.1 in the following example with eight states. But before we come to that, we make some further comments on the average value-at-risk in MDPs which is helpful in the example.

Proposition 8.7.16. *Let $(\Omega,\mathscr{A},P)$ be some probability space. Let X_1, X_2 be random variables with $E|X_1|$, $E|X_2| < \infty$ and $Y \sim \mathrm{Bin}(1,p)$, $0 < p < 1$, where X_1, X_2 and Y are independent. Let $Z \coloneqq \mathbb{1}_{\{0\}}(Y)X_1 + \mathbb{1}_{\{1\}}(Y)X_2$. Then for every $\tau \in (0,1)$*

$$\mathrm{AV@R}_\tau(Z) \geq (1-p)\,\mathrm{AV@R}_\tau(X_1) + p\,\mathrm{AV@R}_\tau(X_2).$$

Proof. Let $\tau \in (0,1)$. From the representation of the average value-at-risk from Theorem 6.2.3, we derive

$$\begin{aligned}
\mathrm{AV@R}_\tau(Z) &\overset{\text{Theorem 6.2.3}}{=} \min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\tau} E\left[(Z-x)^+\right] \right\} \\
&= \min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\tau} \left(P(Y=0)\, E\left[(Z-x)^+ \,\middle|\, Y=0\right] + P(Y=1)\, E\left[(Z-x)^+ \,\middle|\, Y=1\right]\right) \right\} \\
&= \min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\tau} \left((1-p) E\left[(X_1-x)^+ \,\middle|\, Y=0\right] + pE\left[(X_2-x)^+ \,\middle|\, Y=1\right]\right) \right\} \\
&= \min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\tau} \left((1-p) E\left[(X_1-x)^+\right] + pE\left[(X_2-x)^+\right]\right) \right\} \\
&= \min_{x\in\mathbb{R}} \left\{ (1-p) \left(x + \frac{1}{1-\tau} E\left[(X_1-x)^+\right]\right) + p \left(x + \frac{1}{1-\tau} E\left[(X_2-x)^+\right]\right) \right\} \\
&\geq (1-p) \min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\tau} E\left[(X_1-x)^+\right] \right\} + p \min_{x\in\mathbb{R}} \left\{ x + \frac{1}{1-\tau} E\left[(X_2-x)^+\right] \right\} \\
&\overset{\text{Theorem 6.2.3}}{=} (1-p)\,\mathrm{AV@R}_\tau(X_1) + p\,\mathrm{AV@R}_\tau(X_2)
\end{aligned}$$

since X_1, X_2 and Y are independent. □

The preceding result can be generalized to a random variable Y which takes a finite number of values.

Corollary 8.7.17. *Let $(\Omega,\mathscr{A},P)$ be some probability space. Let $X_1, \dots, X_n$ be random variables with $E|X_i| < \infty$, $i = 1, \dots, n$. Let $Y : \Omega \to \{1,\dots,n\}$ be a random variable with distribution $P(Y=i) = p_i$, $i = 1, \dots, n$, and $X_1, \dots, X_n$, Y independent. Let $Z \coloneqq \sum_{i=1}^n \mathbb{1}_{\{i\}}(Y)X_i$. Then for every $\tau \in (0,1)$*

$$\mathrm{AV@R}_\tau(Z) \geq \sum_{i=1}^n p_i\,\mathrm{AV@R}_\tau(X_i).$$

Proof. Essentially, the proof is the same as the proof of Proposition 8.7.16 and is therefore omitted. □

From Corollary 8.7.17, we obtain a lower bound on the average value-at-risk of the average cost for stationary policies.

Proposition 8.7.18. *Let $\tau \in (0,1)$, $x_0 \in S$ and $\mu \in \Pi_s$. Then it holds*

$$\mathrm{AV@R}_\tau^\mu\left(C \,\middle|\, X_0 = x_0\right) \geq \min_{a_0 \in D(x_0)} \sum_{s'\in S} p_{x_0 s'}^{a_0}\,\mathrm{AV@R}_\tau^\mu\left(C \,\middle|\, X_0 = s'\right).$$

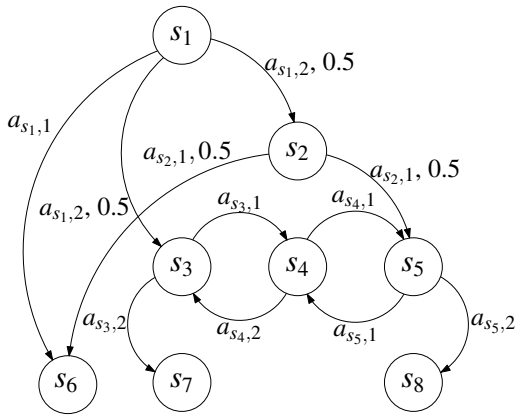


Figure 8.4.: MDP model of Example 8.7.19.

Proof. This is a simple conclusion from Corollary 8.7.17. □

By Proposition 8.7.18, one cannot improve the distribution of the average cost in such a manner that the average value-at-risk of the average cost remains below the optimal average value-at-risk of subsequent states. Now, we come to the numerical example.

Example 8.7.19. Let $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$ and the restriction sets

$$D(s_1) = \{a_{s_1,1}, a_{s_1,2}\}, \quad D(s_2) = \{a_{s_2,1}\}, \quad D(s_3) = \{a_{s_3,1}, a_{s_3,2}\}, \quad D(s_4) = \{a_{s_4,1}, a_{s_4,2}\},$$
$$D(s_5) = \{a_{s_5,1}, a_{s_5,2}\}, \quad D(s_6) = \{a_{s_6,1}\}, \quad D(s_7) = \{a_{s_7,1}\}, \quad D(s_8) = \{a_{s_8,1}\}.$$

The transition probabilities are

$$p_{s_1,s_6}^{a_{s_1,1}} = 1, \quad p_{s_1s_2}^{a_{s_1,2}} = 0.5, \quad p_{s_1,s_3}^{a_{s_1,2}} = 0.5, \quad p_{s_2s_5}^{a_{s_2,1}} = 0.5, \quad p_{s_2s_6}^{a_{s_2,1}} = 0.5, \quad p_{s_3,s_4}^{a_{s_3,1}} = 1, \quad p_{s_3s_7}^{a_{s_3,2}} = 1, \quad p_{s_4s_5}^{a_{s_4,1}} = 1,$$
$$p_{s_4s_3}^{a_{s_4,2}} = 1, \quad p_{s_5s_4}^{a_{s_5,1}} = 1, \quad p_{s_5s_8}^{a_{s_5,2}} = 1, \quad p_{s_6s_6}^{a_{s_6,1}} = 1, \quad p_{s_7s_7}^{a_{s_7,1}} = 1, \quad p_{s_8s_8}^{a_{s_8,1}} = 1$$

and all other transition probabilities are zero. Formally, let $A = \bigcup_{s\in S} D(s)$. The dynamics are shown in Figure 8.4 where the actions of the states s_6, s_7 and s_8 are not depicted. We are looking for an optimal stationary policy with respect to the AV@$R_{0.5}$-criterion, i. e., $\tau = 0.5$. In the following, we omit evident definitions we come across, e. g., when defining policies in the state s_6. Let the initial state be $x_0 = s_1$. The strongly communicating classes are

$$\mathscr{R}_1 = \{s_3, s_4, s_5\}, \quad \mathscr{R}_2 = \{s_6\}, \quad \mathscr{R}_3 = \{s_7\}, \quad \mathscr{R}_4 = \{s_8\}$$

and the class of states which are transient under every policy is $\mathscr{T} = \{s_1, s_2\}$. Furthermore, a stationary policy μ_1 so that $\mathscr{R}_1$ is recurrent is given by

$$\mu_1(s_3, a_{s_3,1}) = 1, \quad \mu_1(s_4, a_{s_4,1}) = 0.5, \quad \mu_1(s_4, a_{s_4,2}) = 0.5, \quad \mu_1(s_5, a_{s_5,1}) = 1.$$

The sinks are the strongly communicating classes with the indices $\mathscr{S} = \{2, 3, 4\}$. For $\mathscr{R}_1$, we have an optimal stationary policy μ_1^* with respect to the expected average cost criterion given by

$$\mu_1^*(s_4, a_{s_4,1}) = \begin{cases} 1, & \text{if } c(s_3, a_{s_3,1}) + c(s_3, a_{s_4,2}) \geq c(s_4, a_{s_4,1}) + c(s_5, a_{s_5,1}) \\ 0, & \text{else} \end{cases},$$
$$\mu_1^*(s_4, a_{s_4,2}) = \begin{cases} 0, & \text{if } c(s_3, a_{s_3,1}) + c(s_3, a_{s_4,2}) \geq c(s_4, a_{s_4,1}) + c(s_5, a_{s_5,1}) \\ 1, & \text{else} \end{cases}.$$

The optimal expected average cost for $\mathscr{R}_1$ are given by $c(1) = \min\{0.5\,c(s_3, a_{s_3,1}) + 0.5\,c(s_4, a_{s_4,2}), 0.5\,c(s_4, a_{s_4,1}) + 0.5\,c(s_5, a_{s_5,1})\}$. Now, we define the MDP $\tilde{\Gamma}$. It is illustrated in Figure 8.5. According to Algorithm 8.1, the next step is to solve MDP $\tilde{\Gamma}$ with respect to the AV@$R_{0.5}$-criterion. At first, following the proof of Theorem 8.7.12, we have a look at all stationary policies so that exactly all states of the set $\tilde{\mathscr{S}} = \{2, 3, 4\}$ are sinks. We are not going to solve the optimization problem (8.13), but we establish an optimal stationary policy by direct computation. Consider the deterministic stationary policies μ^{11}, μ^{12}, μ^{21} and μ^{22} given by

$$\mu^{11}(s_1) = a_{s_1,1}, \quad \mu^{11}(1) = a_{s_3,2},$$

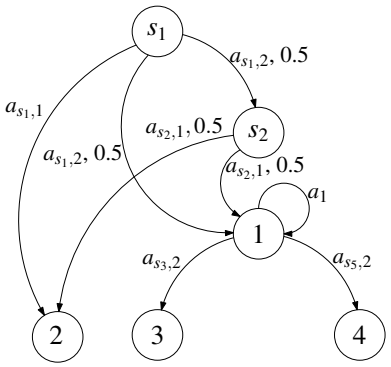


Figure 8.5.: MDP $\tilde{\Gamma}$ in Example 8.7.19.

$$\mu^{12}(s_1) = a_{s_1,1}, \quad \mu^{12}(1) = a_{s_5,2},$$
$$\mu^{21}(s_1) = a_{s_1,2}, \quad \mu^{21}(1) = a_{s_3,2},$$
$$\mu^{22}(s_1) = a_{s_1,2}, \quad \mu^{22}(1) = a_{s_5,2}.$$

We write $c(s_6) = c(s_6, a_{s_6,1})$, $c(s_7) = c(s_7, a_{s_7,1})$, $c(s_8) = c(s_8, a_{s_8,1})$. Then we have

$$\begin{aligned}
&\mathrm{AV@R}_{0.5}^{\mu^{11}}(C \,|\, X_0 = s_1) = \mathrm{AV@R}_{0.5}^{\mu^{12}}(C \,|\, X_0 = s_1) = c(s_6), \\
&\mathrm{AV@R}_{0.5}^{\mu^{21}}(C \,|\, X_0 = s_1) = \left\{ \begin{array}{ll} 0.5\,c(s_6) + 0.5\,c(s_7), & \text{if } c(s_6) \geq c(s_7) \\ c(s_7), & \text{else} \end{array} \right\} = \max\{0.5\,c(s_6) + 0.5\,c(s_7), c(s_7)\}, \\
&\mathrm{AV@R}_{0.5}^{\mu^{22}}(C \,|\, X_0 = s_1) = \left\{ \begin{array}{ll} 0.5\,c(s_6) + 0.5\,c(s_8), & \text{if } c(s_6) \geq c(s_8) \\ c(s_8), & \text{else} \end{array} \right\} = \max\{0.5\,c(s_6) + 0.5\,c(s_8), c(s_8)\}.
\end{aligned}$$

For every randomized stationary policy $\mu = p_{11}\,\mu^{11} + p_{12}\,\mu^{12} + p_{21}\,\mu^{21} + p_{22}\,\mu^{22}$ with $p_{11},\ p_{12},\ p_{21},\ p_{22} \geq 0$ and $p_{11} + p_{12} + p_{21} + p_{22} = 1$, we have

$$\begin{aligned}
&\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = s_1) \\
&\quad \geq p_{11}\,\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = 2) + p_{12}\,\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = 2) + 0.5\,p_{21}\,\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = 1) \\
&\qquad\qquad + 0.5\,p_{21}\,\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = s_2) + 0.5\,p_{22}\,\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = 1) + 0.5\,p_{22}\,\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = s_2) \\
&\quad = (p_{11} + p_{12})\,c(s_6) + p_{21}\max\{0.5\,c(s_6) + 0.5\,c(s_7), c(s_7)\} + p_{22}\max\{0.5\,c(s_6) + 0.5\,c(s_8), c(s_8)\} \qquad (8.14)
\end{aligned}$$

by Corollary 8.7.17. The term (8.14) is minimized if

$$\begin{aligned}
&p_{11} = 1, \text{ in the case } c(s_6) \leq c(s_7) \leq c(s_8) \text{ or } c(s_6) \leq c(s_8) \leq c(s_7), \\
&p_{21} = 1, \text{ in the case } c(s_7) \leq c(s_6) \leq c(s_8) \text{ or } c(s_7) \leq c(s_8) \leq c(s_6), \\
&p_{22} = 1, \text{ in the case } c(s_8) \leq c(s_6) \leq c(s_7) \text{ or } c(s_8) \leq c(s_7) \leq c(s_6).
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = s_1) \geq c(s_6), \text{ if } c(s_6) \leq c(s_7) \leq c(s_8) \text{ or } c(s_6) \leq c(s_8) \leq c(s_7), \\
&\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = s_1) \geq \max\{0.5\,c(s_6) + 0.5\,c(s_7), c(s_7)\}, \text{ if } c(s_7) \leq c(s_6) \leq c(s_8) \text{ or } c(s_7) \leq c(s_8) \leq c(s_6), \\
&\mathrm{AV@R}_{0.5}^{\mu}(C \,|\, X_0 = s_1) \geq \max\{0.5\,c(s_6) + 0.5\,c(s_8), c(s_8)\}, \text{ if } c(s_8) \leq c(s_6) \leq c(s_7) \text{ or } c(s_8) \leq c(s_7) \leq c(s_6),
\end{aligned}$$

where equality holds for μ^{11}, μ^{21} and μ^{22} respectively. Hence, depending on the costs $c(s_6)$, $c(s_7)$ and $c(s_8)$, one of the deterministic policies μ^{11}, μ^{21} and μ^{22} is optimal. Now, we consider the case where the indices of the sinks are $\tilde{\mathscr{S}} = \{1,2,3,4\}$. In this case, we need to consider just two stationary policies μ^1 and μ^2 with

$$\mu^1(s_1) = a_{s_1,1} \quad \text{and} \quad \mu^2(s_1) = a_{s_1,2}.$$

In this case, we have

$$\mathrm{AV@R}_{0.5}^{\mu^1}(C \,|\, X_0 = s_1) = c(s_6),$$

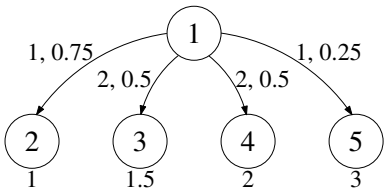


Figure 8.6.: MDP model of Example 8.8.1.

$$\mathrm{AV@R}_{0.5}^{\mu^2}(C\,|\,X_0 = s_1) = \max\{c(s_6), c(1)\}.$$

The same technique as before establishes optimality of μ^1 in the case $c(s_6) \leq c(1)$ and optimality of μ^2 in the case $c(s_6) \geq c(1)$. Hence, we can give deterministic $\mathrm{AV@R}_{0.5}$-optimal stationary policies for the MDP $\tilde{\Gamma}$. Note that the cost $c(1)$ does not play any role since μ^{11} performs at least as good as μ^1 and μ^2 in any case. The $\mathrm{AV@R}_{0.5}$-optimal stationary policies for $\tilde{\Gamma}$ are

$$\begin{aligned}
&\mu^{11}, \text{ in the case } c(s_6) \leq c(s_7) \leq c(s_8) \text{ or } c(s_6) \leq c(s_8) \leq c(s_7),\\
&\mu^{21}, \text{ in the case } c(s_7) \leq c(s_6) \leq c(s_8) \text{ or } c(s_7) \leq c(s_8) \leq c(s_6),\\
&\mu^{22}, \text{ in the case } c(s_8) \leq c(s_6) \leq c(s_7) \text{ or } c(s_8) \leq c(s_7) \leq c(s_6).
\end{aligned}$$

Now, we are able to put together the results and specify an over-all optimal policy. From now on, let $c(s_7) \leq c(s_6) \leq c(s_8)$, for example. In this case, an optimal stationary policy μ^* is given by

$$\begin{aligned}
&\mu^*(s_1) = a_{s_1,2}, \quad \mu^*(s_2) = a_{s_2,1}, \quad \mu^*(s_3) = a_{s_3,2}, \quad \mu_1(s_4, a_{s_4,1}) = 0.5, \quad \mu_1(s_4, a_{s_4,2}) = 0.5, \quad \mu_1(s_5, a_{s_5,1}) = 1,\\
&\mu^*(s_5) = a_{s_5,1},
\end{aligned}$$

which leads to the optimal value $\mathrm{AV@R}_{0.5}^{\mu^*}(C\,|\,X_0 = x_0) = c(s_7)$. Indeed, defining $\mu^*(s_4) = a_{s_4,2}$ leads to a deterministic optimal stationary policy.

8.8. A Remark on Average Value-at-Risk-Optimal Policies

As already mentioned in Example 8.7.6, the function $\mu \mapsto \mathrm{AV@R}_\tau^\mu(C\,|\,X_0 = x_0)$ is not continuous according to any vector norm on Π_s. Another remark is that a convex combination of $\mathrm{AV@R}_\tau$-optimal stationary policies need not be an $\mathrm{AV@R}_\tau$-optimal stationary policy in general. This is different to the case of the expected cost criterion where the convex combination of two optimal policies is optimal, too. This is shown in the following example.

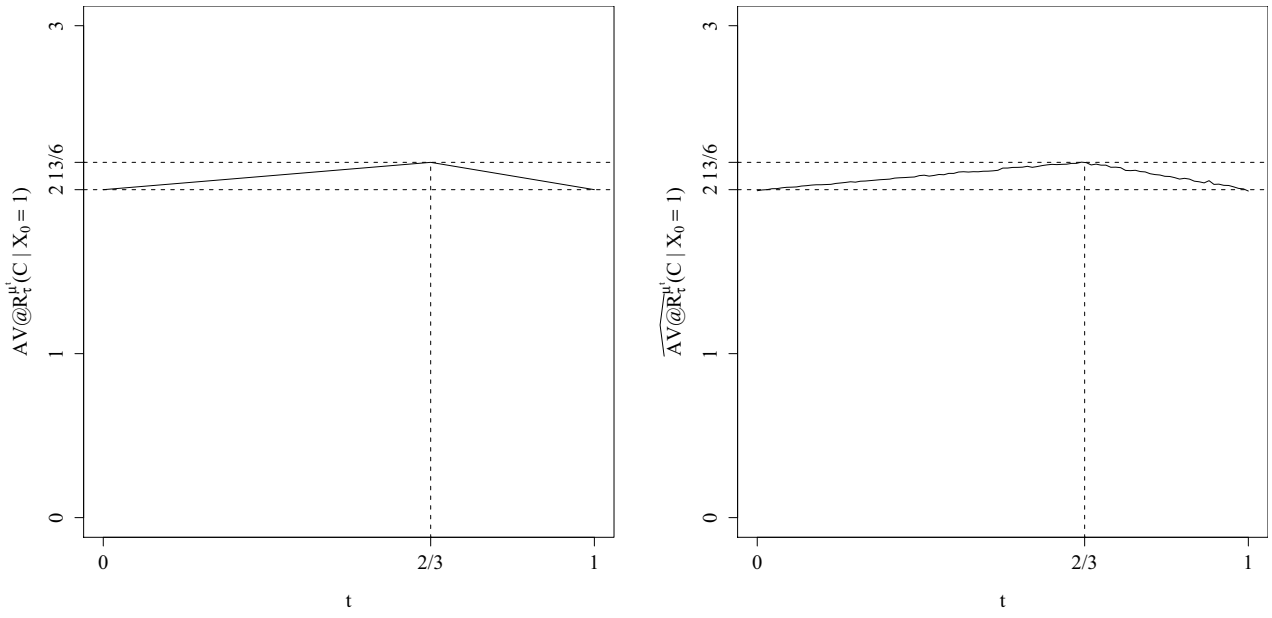
Example 8.8.1. Let the state space be $S = \{1,2,3,4,5\}$, the action space be $A = \{1,2\}$ and the restriction sets be $D(1) = A$ and $D(s) = \{1\}$, $s = 2, 3, 4, 5$. The transition probabilities and the costs are given by

$$\begin{aligned}
&p_{12}^1 = 0.75, \quad p_{15}^1 = 0.25, \quad p_{13}^2 = 0.5, \quad p_{14}^2 = 0.5,\\
&c(1) = 0, \quad c(2) = 1, \quad c(3) = 1.5, \quad c(4) = 2, \quad c(5) = 3
\end{aligned}$$

and $p_{ss}^1 = 1$, $s = 2, 3, 4, 5$, where the remaining transition probabilities are zero. The model is depicted in Figure 8.6 where trivial actions which lead from one state to the same with probability one are blanked out. The decision maker has only to decide which action should be chosen in state 1. Let the confidence level be $\tau = 0.5$. Then the decision rules μ_1 and μ_2 defined by $\mu_1(1|1) = 1$ and $\mu_2(2|1) = 1$ respectively lead to the following values of the average value-at-risk of the average cost:

$$\begin{aligned}
\mathrm{AV@R}_\tau^{\mu_1}(C\,|\,X_0 = 1) &= \frac{0.25 \cdot 3 + 0.25 \cdot 1}{0.5} = 2,\\
\mathrm{AV@R}_\tau^{\mu_2}(C\,|\,X_0 = 1) &= \frac{0.5 \cdot 2}{0.5} = 2.
\end{aligned}$$

For $t \in [0,1]$, define the stationary policy μ^t by $\mu^t := t\,\mu_1 + (1-t)\,\mu_2$. In the next step, we want to calculate $\mathrm{AV@R}_\tau^{\mu^t}(C\,|\,X_0 = 1)$ for all $t \in [0,1]$. We have $\mathrm{AV@R}_\tau^{\mu^t}(C\,|\,X_0 = 1) \geq t\,\mathrm{AV@R}_\tau^{\mu_1}(C\,|\,X_0 = 1) + (1-t)\,\mathrm{AV@R}_\tau^{\mu_2}(C\,|\,X_0 = 1) = 2$ for every $t \in [0,1]$ by Proposition 8.7.16. Next, we show that indeed strict inequality holds for every $t \in (0,1)$. To begin



(a) Function $t \mapsto \text{AV@R}_\tau^{\mu^t}(C \,|\, X_0 = 1)$.

(b) Simulation of $\text{AV@R}_\tau^{\mu^t}(C \,|\, X_0 = 1)$.

Figure 8.7.: Results of Example 8.8.1.

with the calculation, note that for no $t \in [0,1]$ we have $P^{\mu^t}(C = 3 \,|\, X_0 = 1) \geq 0.5$. Hence $\text{AV@R}_\tau^{\mu^t}(C \,|\, X_0 = 1) < 3$ for all $t \in [0,1]$. Further, we have

$$P^{\mu^t}(C \geq 2 \,|\, X_0 = 1) \geq 0.5 \quad \Leftrightarrow \quad 0.25t + 0.5(1-t) \geq 0.5 \quad \Leftrightarrow \quad t \leq 0 \quad \Leftrightarrow \quad t = 0.$$

In this case, $\text{AV@R}_\tau^{\mu^0}(C \,|\, X_0 = 1) = 2$ as computed earlier. We precede by calculating those $t \in [0,1]$, for which

$$P^{\mu^t}(C \geq 1.5 \,|\, X_0 = 1) \geq 0.5 \quad \Leftrightarrow \quad 0.25t + 0.5(1-t) + 0.5(1-t) \geq 0.5 \quad \Leftrightarrow \quad t \leq \frac{2}{3}.$$

For $t \in (0, 2/3]$, we then have

$$\text{AV@R}_\tau^{\mu^t}(C \,|\, X_0 = 1) = \frac{3 \cdot 0.25t + 2 \cdot 0.5(1-t) + 1.5 \cdot (0.5 - 0.25t - 0.5(1-t))}{0.5} = 2 + 0.25t.$$

In the last case $t \in (2/3, 1]$, we have

$$\begin{aligned}\text{AV@R}_\tau^{\mu^t}(C \,|\, X_0 = 1) &= \frac{3 \cdot 0.25t + 2 \cdot 0.5(1-t) + 1.5 \cdot 0.5(1-t) + 1 \cdot (0.5 - 0.25t - 0.5(1-t) - 0.5(1-t))}{0.5} \\ &= 2.5 - 0.5t.\end{aligned}$$

Hence, μ_1 and μ_2 both are AV@R_τ-optimal, whereas any strictly convex combination of these policies performs worse with respect to the $\text{AV@R}_{0.5}$-criterion. The maximum value of $\text{AV@R}_{0.5}^{\mu^t}(C \,|\, X_0 = 1)$ is reached for $t = 2/3$ with $\text{AV@R}_{0.5}^{\mu^{2/3}}(C \,|\, X_0 = 1) = 13/6$. In Figure 8.7, these results are illustrated as well as a an estimate of $\text{AV@R}_\tau^{\mu^t}$ from a simulation, from which the same facts are apparent.

A. Miscellaneous Lemmas

In this chapter, some lemmas which are used in the prior text are stated and proved.

Lemma A.1. *The positive part function* $(\cdot)^+ : \mathbb{R} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^+ = \max\{0, x\}$, *is increasing and convex.*

Proof. These statements are obvious. □

Lemma A.2. *The positive part function is positively homogeneous, i. e., for every* $x \in \mathbb{R}$ *and* $c \geq 0$, *it holds*

$$(cx)^+ = cx^+.$$

Proof. Let $x \in \mathbb{R}$ and $c \geq 0$. By definition of the positive part function and non-negativity of c, one computes

$$(cx)^+ = \max\{0, cx\} = c\max\{0, x\} = cx^+.$$ □

Lemma A.3. *Let* $a, b \in \mathbb{R}$.

1. *If* $b \geq 0$, *then*

$$(a+b)^+ \leq a^+ + b.$$

2. *If* $b \leq 0$, *then*

$$(a+b)^+ \geq a^+ + b.$$

Proof. For 1, easy computation yields

$$a^+ + b = \max\{0, a\} + b = \max\{b, a+b\} \geq \{0, a+b\} = (a+b)^+$$

since $b \geq 0$. For $b \leq 0$, we have the transition functions

$$a^+ + b = \max\{0, a\} + b = \max\{b, a+b\} \leq \{0, a+b\} = (a+b)^+,$$

yielding 2. □

Lemma A.4. *Let* $X, Y \neq \emptyset$ *be arbitrary sets and* $f : X \times Y \to \mathbb{R}$. *Then it holds*

$$\inf_{x \in X} \inf_{y \in Y} f(x, y) = \inf_{y \in Y} \inf_{x \in X} f(x, y).$$

Proof. Let $x' \in X$. Then we have $f(x', y) \geq \inf_{x \in X} f(x, y)$ for all $y \in Y$. Hence $\inf_{y \in Y} f(x', y) \geq \inf_{y \in Y} \inf_{x \in X} f(x, y)$. Since x' is arbitrary, we have $\inf_{x \in X} \inf_{y \in Y} f(x', y) \geq \inf_{y \in Y} \inf_{x \in X} f(x, y)$. For the reverse inequality, take some $y' \in Y$. Then $\inf_{y \in Y} f(x, y) \leq f(x, y')$ for all $x \in X$. Hence $\inf_{x \in X} \inf_{y \in Y} f(x, y) \leq \inf_{x \in X} f(x, y')$, and finally, it follows $\inf_{x \in X} \inf_{y \in Y} f(x, y) \leq \inf_{y \in Y} \inf_{x \in X} f(x, y)$. □

Lemma A.5. *Let* $X \neq \emptyset$ *be an arbitrary set. Further, let* f, $g : X \to \mathbb{R}$ *be real-valued functions, where* f *or* g *is bounded below. Then it holds*

$$\left| \inf_{x \in X} f(x) - \inf_{x \in X} g(x) \right| \leq \sup_{x \in X} |f(x) - g(x)|.$$

Proof. At first, let f and g be bounded below. Further, let $\varepsilon > 0$ arbitrary and fixed. Then from the definition of the infimum, there is some $x_1 \in X$ such that $f(x_1) - \varepsilon \leq \inf_{x\in X} f(x)$. Similarly, there exists some $x_2 \in X$ such that $g(x_2) - \varepsilon \leq \inf_{x\in X} g(x)$. With these x_1 and x_2, one computes

$$\begin{aligned}
\left|\inf_{x\in X} f(x) - \inf_{x\in X} g(x)\right| &= \begin{cases} \inf_{x\in X} f(x) - \inf_{x\in X} g(x), & \text{if } \inf_{x\in X} f(x) \geq \inf_{x\in X} g(x) \\ \inf_{x\in X} g(x) - \inf_{x\in X} f(x), & \text{if } \inf_{x\in X} f(x) < \inf_{x\in X} g(x) \end{cases} \\
&\leq \left\{ \begin{array}{ll} f(x_2) - g(x_2) + \varepsilon, & \text{if } \inf_{x\in X} f(x) \geq \inf_{x\in X} g(x) \\ g(x_1) - f(x_1) + \varepsilon, & \text{if } \inf_{x\in X} f(x) < \inf_{x\in X} g(x) \end{array} \right\} \qquad \text{(A.1)} \\
&\leq \begin{cases} |f(x_2) - g(x_2)| + \varepsilon, & \text{if } \inf_{x\in X} f(x) \geq \inf_{x\in X} g(x) \\ |f(x_1) - g(x_1)| + \varepsilon, & \text{if } \inf_{x\in X} f(x) < \inf_{x\in X} g(x) \end{cases} \\
&\leq \sup_{x\in X} |f(x) - g(x)| + \varepsilon. \qquad \text{(A.2)}
\end{aligned}$$

Since (A.2) holds for every $\varepsilon > 0$, we get $|\inf_{x\in X} f(x) - \inf_{x\in X} g(x)| \leq \sup_{x\in X} |f(x) - g(x)|$. In the remaining case, let f be bounded below and g be unbounded below w. l. o. g.. Then $|\inf_{x\in X} f(x) - \inf_{x\in X} g(x)| = \infty$ and $\sup_{x\in X} |f(x) - g(x)| = \infty$. The assertion holds in this case, too. □

Lemma A.6. *Let f, $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constants L_f and L_g respectively. Then $f + g$ and cf for some $c \in \mathbb{R}$ are Lipschitz continuous with Lipschitz constants $L_f + L_g$ and $|c| L_f$ respectively.*

Proof. Let $x, y \in \mathbb{R}$. Then

$$\begin{aligned}
&|(f+g)(x) - (f+g)(y)| \leq |f(x) - f(y)| + |g(x) - g(y)| \leq (L_f + L_g)\,|x - y| \quad \text{and} \\
&|(cf)(x) - (cf)(y)| = |c|\,|f(x) - f(y)| \leq |c| L_f\,|x - y|.
\end{aligned}$$

□

Lemma A.7. *Let $f_n : \mathbb{R} \to \mathbb{R}$, $n \in \mathbb{N}_0$, be decreasing with $f_n(x) \overset{n\to\infty}{\to} f(x)$ for all $x \in \mathbb{R}$. Then f is decreasing.*

Proof. Let $\varepsilon > 0$ and $x \leq y$. Then there is an $n_0 \in \mathbb{N}_0$ such that $|f_n(x) - f(x)| < \varepsilon/2$ and $|f_n(y) - f(y)| < \varepsilon/2$ for all $n \geq n_0$. Therefore

$$f(y) - f(x) = \underbrace{f(y) - f_n(y)}_{\in(-\varepsilon/2,\varepsilon/2)} - \underbrace{(f(x) - f_n(x))}_{\in(-\varepsilon/2,\varepsilon/2)} + \underbrace{f_n(y) - f_n(x)}_{\leq 0} \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude $f(x) \geq f(y)$. □

B. Parameters of the Numerical Examples

For completeness, the parameters of the numerical examples that are considered in the above chapters are listed in this chapter.

B.1. Parameters of Example 4.1.8

The infrastucture consists of a single sector $\Sigma = \{\sigma\}$. The state space is $G = \{0,1,\dots,4\}$. The elementary actions are $A_0 = \{0,2\}$. Action 2 satisfies Assumption 4.1.5.4. We have three threat events $\mathscr{E} = \{e_1, e_2, e_3\}$ with

$$\Psi_{e_1}(g) = 0, \quad \Psi_{e_2}(g) = \min\{g+2,4\}, \quad \Psi_{e_3}(g) = \begin{cases} 1, & \text{if } g = 0,1,2 \\ g+1, & \text{else} \end{cases}, \quad g \in G,$$
$$C_{e_1} = 1000, \quad C_{e_2} = 0, \quad C_{e_3} = 0,$$
$$\lambda_{e_1}(0) = \frac{1}{8760}, \quad \lambda_{e_2}(0) = \frac{1}{4380}, \quad \lambda_{e_3}(0) = \frac{1}{4380},$$
$$c(0) = 1, \quad c(1) = \frac{365}{7}, \quad c(2) = 365, \quad c(3) = 8760, \quad c(4) = 17520.$$

The discount rate is set to $\alpha = 1/1000$.

B.2. Parameters of the numerical example in section 5.4.7

The infrastructure consists of five sectors such that $\Sigma = \{\sigma_1, \dots, \sigma_5\}$. The dependency structure is given by

$$N = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

We have five threat levels so that $G = \{0,\dots,4\}$. For every sector, we have three threat events $\mathscr{E}(\sigma_i) = \{e^1_{\sigma_i}, e^2_{\sigma_i}, e^3_{\sigma_i}\}$, $i = 1,\dots,5$, where the first threat event models the destruction, the second one models an alarm and the third one models a "Nothing happened" event of the respective sector. We let

$$\lambda_{e^1_{\sigma_i}}(0) = \frac{1}{876{,}000}, \quad \lambda_{e^1_{\sigma_i}}(1) = \frac{1}{87{,}600}, \quad \lambda_{e^1_{\sigma_i}}(2) = \frac{1}{8{,}760}, \quad \lambda_{e^1_{\sigma_i}}(3) = \frac{1}{730}, \quad \lambda_{e^1_{\sigma_i}}(4) = \frac{2}{24},$$
$$\lambda_{e^2_{\sigma_i}}(0) = \frac{2}{8{,}760}, \quad \lambda_{e^2_{\sigma_i}}(1) = \frac{1}{730}, \quad \lambda_{e^2_{\sigma_i}}(2) = \frac{1}{48}, \quad \lambda_{e^2_{\sigma_i}}(3) = \frac{2}{3}, \quad \lambda_{e^2_{\sigma_i}}(4) = 40,$$
$$\lambda_{e^3_{\sigma_i}}(0) = 4, \quad \lambda_{e^3_{\sigma_i}}(1) = 3, \quad \lambda_{e^3_{\sigma_i}}(2) = 2, \quad \lambda_{e^3_{\sigma_i}}(3) = 1, \quad \lambda_{e^3_{\sigma_i}}(4) = 0, \quad i = 1,\dots,5.$$

The costs are

$$C_{e^1_{\sigma_1}} = 119{,}680, \quad C_{e^1_{\sigma_2}} = 746{,}400, \quad C_{e^1_{\sigma_3}} = 418{,}000, \quad C_{e^1_{\sigma_4}} = 488{,}000, \quad C_{e^1_{\sigma_5}} = 460{,}000,$$
$$C_{e^2_{\sigma_1}} = 1, \quad C_{e^2_{\sigma_2}} = 1, \quad C_{e^2_{\sigma_3}} = 1, \quad C_{e^2_{\sigma_4}} = 1, \quad C_{e^2_{\sigma_5}} = 1,$$
$$C_{e^3_{\sigma_1}} = 0, \quad C_{e^3_{\sigma_2}} = 0, \quad C_{e^3_{\sigma_3}} = 0, \quad C_{e^3_{\sigma_4}} = 0, \quad C_{e^3_{\sigma_5}} = 0.$$

The transition functions for the affected sectors are

$$\Psi_{e^1_{\sigma_i}}(g) = 0,$$
$$\Psi_{e^2_{\sigma_i}}(g) = \min\{g+2,4\},$$

$$\Psi_{e^3_{\sigma_i}}(g) = \begin{cases} 1, & \text{if } g = 0,1 \\ g-1, & \text{else} \end{cases}, \quad g \in G, \quad i = 1,\dots,5.$$

For dependent sectors, we have

$$\begin{aligned}
&\psi_{e^1_{\sigma_i}}(g) = 4,\\
&\psi_{e^2_{\sigma_i}}(g) = \min\{g+1,4\},\\
&\psi_{e^3_{\sigma_i}}(g) = g, \quad g \in G, \quad i = 1,\dots,5.
\end{aligned}$$

Further, we have two active elementary actions: elementary action 1 models a camera evaluation and 2 models an inspection walk. The rates are

$$\lambda_1(\sigma_i) = 30, \quad \lambda_2(\sigma_i) = 10, \quad i = 1,\dots,5,$$

and the transition probabilities for the affected sector are given by

$$\begin{aligned}
&\Phi^0_{1,\sigma_i}(0) = 0.99, \quad \Phi^0_{1,\sigma_i}(1) = 0, \quad \Phi^0_{1,\sigma_i}(2) = 0, \quad \Phi^0_{1,\sigma_i}(3) = 0, \quad \Phi^0_{1,\sigma_i}(4) = 0.01,\\
&\Phi^1_{1,\sigma_i}(0) = 0.8, \quad \Phi^1_{1,\sigma_i}(1) = 0, \quad \Phi^1_{1,\sigma_i}(2) = 0, \quad \Phi^1_{1,\sigma_i}(3) = 0, \quad \Phi^1_{1,\sigma_i}(4) = 0.2,\\
&\Phi^2_{1,\sigma_i}(0) = 0.5, \quad \Phi^2_{1,\sigma_i}(1) = 0, \quad \Phi^2_{1,\sigma_i}(2) = 0, \quad \Phi^2_{1,\sigma_i}(3) = 0, \quad \Phi^2_{1,\sigma_i}(4) = 0.5,\\
&\Phi^3_{1,\sigma_i}(0) = 0.2, \quad \Phi^3_{1,\sigma_i}(1) = 0, \quad \Phi^3_{1,\sigma_i}(2) = 0, \quad \Phi^3_{1,\sigma_i}(3) = 0, \quad \Phi^3_{1,\sigma_i}(4) = 0.8,\\
&\Phi^4_{1,\sigma_i}(0) = 0.01, \quad \Phi^4_{1,\sigma_i}(1) = 0, \quad \Phi^4_{1,\sigma_i}(2) = 0, \quad \Phi^4_{1,\sigma_i}(3) = 0, \quad \Phi^4_{1,\sigma_i}(4) = 0.99,\\
&\Phi^0_{2,\sigma_i}(0) = 1, \quad \Phi^0_{2,\sigma_i}(1) = 0, \quad \Phi^0_{2,\sigma_i}(2) = 0, \quad \Phi^0_{2,\sigma_i}(3) = 0, \quad \Phi^0_{2,\sigma_i}(4) = 0,\\
&\Phi^1_{2,\sigma_i}(0) = 1, \quad \Phi^1_{2,\sigma_i}(1) = 0, \quad \Phi^1_{2,\sigma_i}(2) = 0, \quad \Phi^1_{2,\sigma_i}(3) = 0, \quad \Phi^1_{2,\sigma_i}(4) = 0,\\
&\Phi^2_{2,\sigma_i}(0) = 1, \quad \Phi^2_{2,\sigma_i}(1) = 0, \quad \Phi^2_{2,\sigma_i}(2) = 0, \quad \Phi^2_{2,\sigma_i}(3) = 0, \quad \Phi^2_{2,\sigma_i}(4) = 0,\\
&\Phi^3_{2,\sigma_i}(0) = 1, \quad \Phi^3_{2,\sigma_i}(1) = 0, \quad \Phi^3_{2,\sigma_i}(2) = 0, \quad \Phi^3_{2,\sigma_i}(3) = 0, \quad \Phi^3_{2,\sigma_i}(4) = 0,\\
&\Phi^4_{2,\sigma_i}(0) = 1, \quad \Phi^4_{2,\sigma_i}(1) = 0, \quad \Phi^4_{2,\sigma_i}(2) = 0, \quad \Phi^4_{2,\sigma_i}(3) = 0, \quad \Phi^4_{2,\sigma_i}(4) = 0, \quad i = 1,\dots,5.
\end{aligned}$$

The transition functions for the dependent sectors are

$$\varphi_1(g,g',g^*) = \begin{cases} \min\left\{0, \max\left\{4, g^* + \lfloor \frac{g'-g}{2} \rfloor\right\}\right\}, & \text{if } g^* = 0 \\ \min\left\{1, \max\left\{4, g^* + \lfloor \frac{g'-g}{2} \rfloor\right\}\right\}, & \text{if } g^* \neq 0 \end{cases},$$
$$\varphi_2(g,g',g^*) = 4, \quad g, g', g^* \in G.$$

The elementary actions do not cost anything, i. e., $c_1 = c_2 = 0$. We have $\gamma_{\sigma_i} \equiv 0$ and $C_{\sigma_i} = 0$, $i = 1,\dots,5$. The discount rate is $\alpha = \log(2)/24$.

B.3. Parameters of the numerical example in section 7.5.3

In this section, we present the parameters of the example in section 7.5.3. The state space is $S := \{1,2,3\}$ and the action space $A = \{1,2,3\}$. The restriction set is $D(s) := A$, $s \in S$. The costs take values in $W := \{0,1,2,3\}$. The discount factor is $\beta := 0.8$. The transition probabilities are given by

$$\begin{aligned}
&p^1_{10} = 0.1, \quad p^1_{11} = 0.4, \quad p^1_{12} = 0.2, \quad p^1_{13} = 0.3,\\
&p^2_{10} = 0.4, \quad p^2_{11} = 0.0, \quad p^2_{12} = 0.3, \quad p^2_{13} = 0.3,\\
&p^3_{10} = 0.1, \quad p^3_{11} = 0.2, \quad p^3_{12} = 0.6, \quad p^3_{13} = 0.1,\\
&p^1_{20} = 0.6, \quad p^1_{21} = 0.3, \quad p^1_{22} = 0.1, \quad p^1_{23} = 0.0,\\
&p^2_{20} = 0.1, \quad p^2_{21} = 0.1, \quad p^2_{22} = 0.3, \quad p^2_{23} = 0.5,\\
&p^3_{20} = 0.6, \quad p^3_{21} = 0.2, \quad p^3_{22} = 0.1, \quad p^3_{23} = 0.1,\\
&p^1_{30} = 0.3, \quad p^1_{31} = 0.2, \quad p^1_{32} = 0.1, \quad p^1_{33} = 0.4,\\
&p^2_{30} = 0.0, \quad p^2_{31} = 0.3, \quad p^2_{32} = 0.5, \quad p^2_{33} = 0.2,
\end{aligned}$$

$p^3_{30} = 0.3, \quad p^3_{31} = 0.3, \quad p^3_{32} = 0.2, \quad p^3_{33} = 0.2,$
$p^1_{110} = 0.2, \quad p^1_{111} = 0.3, \quad p^1_{112} = 0.0, \quad p^1_{113} = 0.0,$
$p^1_{120} = 0.0, \quad p^1_{121} = 0.2, \quad p^1_{122} = 0.0, \quad p^1_{123} = 0.1,$
$p^1_{130} = 0.0, \quad p^1_{131} = 0.1, \quad p^1_{132} = 0.1, \quad p^1_{133} = 0.0,$
$p^2_{110} = 0.0, \quad p^2_{111} = 0.0, \quad p^2_{112} = 0.1, \quad p^2_{113} = 0.1,$
$p^2_{120} = 0.0, \quad p^2_{121} = 0.2, \quad p^2_{122} = 0.1, \quad p^2_{123} = 0.0,$
$p^2_{130} = 0.1, \quad p^2_{131} = 0.2, \quad p^2_{132} = 0.1, \quad p^2_{133} = 0.1,$
$p^3_{110} = 0.2, \quad p^3_{111} = 0.0, \quad p^3_{112} = 0.0, \quad p^3_{113} = 0.0,$
$p^3_{120} = 0.5, \quad p^3_{121} = 0.0, \quad p^3_{122} = 0.0, \quad p^3_{123} = 0.0,$
$p^3_{130} = 0.3, \quad p^3_{131} = 0.0, \quad p^3_{132} = 0.0, \quad p^3_{133} = 0.0,$
$p^1_{210} = 0.1, \quad p^1_{211} = 0.1, \quad p^1_{212} = 0.1, \quad p^1_{213} = 0.0,$
$p^1_{220} = 0.0, \quad p^1_{221} = 0.0, \quad p^1_{222} = 0.1, \quad p^1_{223} = 0.2,$
$p^1_{230} = 0.3, \quad p^1_{231} = 0.0, \quad p^1_{232} = 0.0, \quad p^1_{233} = 0.1,$
$p^2_{210} = 0.1, \quad p^2_{211} = 0.0, \quad p^2_{212} = 0.0, \quad p^2_{213} = 0.3,$
$p^2_{220} = 0.0, \quad p^2_{221} = 0.0, \quad p^2_{222} = 0.0, \quad p^2_{223} = 0.3,$
$p^2_{230} = 0.0, \quad p^2_{231} = 0.0, \quad p^2_{232} = 0.0, \quad p^2_{233} = 0.3,$
$p^3_{210} = 0.1, \quad p^3_{211} = 0.0, \quad p^3_{212} = 0.0, \quad p^3_{213} = 0.2,$
$p^3_{220} = 0.0, \quad p^3_{221} = 0.0, \quad p^3_{222} = 0.2, \quad p^3_{223} = 0.3,$
$p^3_{230} = 0.2, \quad p^3_{231} = 0.0, \quad p^3_{232} = 0.0, \quad p^3_{233} = 0.0,$
$p^1_{310} = 0.0, \quad p^1_{311} = 0.0, \quad p^1_{312} = 0.0, \quad p^1_{313} = 0.0,$
$p^1_{320} = 0.3, \quad p^1_{321} = 0.0, \quad p^1_{322} = 0.0, \quad p^1_{323} = 0.4,$
$p^1_{330} = 0.0, \quad p^1_{331} = 0.1, \quad p^1_{332} = 0.2, \quad p^1_{333} = 0.0,$
$p^2_{310} = 0.1, \quad p^2_{311} = 0.3, \quad p^2_{312} = 0.2, \quad p^2_{313} = 0.0,$
$p^2_{320} = 0.0, \quad p^2_{321} = 0.0, \quad p^2_{322} = 0.0, \quad p^2_{323} = 0.2,$
$p^2_{330} = 0.0, \quad p^2_{331} = 0.0, \quad p^2_{332} = 0.0, \quad p^2_{333} = 0.2,$
$p^3_{310} = 0.0, \quad p^3_{311} = 0.1, \quad p^3_{312} = 0.2, \quad p^3_{313} = 0.0,$
$p^3_{320} = 0.0, \quad p^3_{321} = 0.2, \quad p^3_{322} = 0.2, \quad p^3_{323} = 0.0,$
$p^3_{330} = 0.0, \quad p^3_{331} = 0.2, \quad p^3_{332} = 0.1, \quad p^3_{333} = 0.0.$

C. Mathematical Symbols and Notation

$\mathbb{N} = \{1, 2, \ldots\}$	set of natural numbers
$\mathbb{N}_0 = \{0, 1, 2, \ldots\}$	set of natural numbers including 0
$\mathbb{R}$	set of real numbers
$\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\} = [0, \infty)$	set of non-negative real numbers
$\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\} = (0, \infty)$	set of positive real numbers
S	state space
A	action space
D	restriction set
$D(s)$	admissible actions in state s
W	set of costs
$\lambda(s,a)$	rate until transition when the current state is s under action a
$p_{ss'}^a$	transition probability from state s to s' under action a
$p_{ss'c}^a$	transition probability from state s to s' and incurring cost c under action a
α	discount rate
β	discount factor
$\mathscr{K}$	setup cost structure
$\mathscr{C}$	cost structure
$c(s,a)$	cost of being in state s and taking action a (for MDPs) or the expected discounted one-step cost of being in state s and taking action a (for CMDPs)
Π	set of randomized history-dependent policies
Π^d	set of deterministic history-dependent policies
Π_m	set of randomized Markovian policies
Π_m^d	set deterministic Markovian policies
Π_s	set of randomized stationary policies
Π_s^d	set of deterministic stationary policies
π	arbitrary policy
π^*	optimal policy
μ	stationary policy
μ^*	stationary optimal policy or optimal stationary policy in chapter 8
P^π	distribution under policy π
E^π	expectation under policy π
$(X_k)_{k\in\mathbb{N}_0}$, $(\tilde{X}_t)_{t\geq 0}$	state process
$(A_k)_{k\in\mathbb{N}_0}$, $(\tilde{A}_t)_{t\geq 0}$	action process
$(C_k)_{k\in\mathbb{N}_0}$	cost process
$\lvert M\rvert$	number of elements of the set M
$\mathscr{P}(M)$	power set of the set M
$\lVert v\rVert_\infty = \sup_{x\in M} \lvert v(x)\rvert$	supremum norm for $v : M \to \mathbb{R}$
$\mathbb{1}_A$	indicator function of the set A
Σ	set of sectors of the infrastructure
N	dependency matrix
$G = \{0, \ldots, g_{\max}\}$	set of threat levels
$\mathscr{E}(\sigma)$	set of threat events of sector σ
Ψ_e	transition function for the affected sector when threat event e occurs
ψ_e	transition function for dependent sectors when threat event e occurs
C_e	cost of threat event e
$\lambda_e(g)$	rate until occurrence of threat event e at threat level g
$\gamma_\sigma(g)$	probability of finding a dangerous object in sector σ at threat level g
C_σ	cost of removing dangerous object from sector σ

A_0	set of elementary actions
$\Phi_{a_0,\sigma}$	transition mechanism when elementary action a_0 is accomplished in sector σ
φ_{a_0}	transition function for dependent sectors when elementary action a_0 is accomplished
$\lambda_\sigma(a_0)$	rate of accomplishing elementary action a_0 in sector σ
c_{a_0}	cost rate of elementary action a_0
κ	coupling mechanism probability
P^{c}	coupling transition probabilities
$\imath, \tilde{\imath}$	heuristic indices
C^n	total discounted cost over a finite horizon of length n
C^∞	total discounted cost over an infinite horizon
C_-	lim inf of the average cost
C_+	lim sup of the average cost
C	long-run average cost
ρ	risk measure
τ	confidence level
$\mathrm{AV@R}_\tau^\pi(X)$	average value-at-risk at level τ of X under policy π
$\mathrm{V@R}_\tau^\pi(X)$	value-at-risk at level τ of X under policy π
$\alpha(\cdot,\cdot)$	absorption probabilities
v_n^π	n-horizon value function under policy π
v_n^*	optimal n-horizon value function
v^π	value function under policy π
v^*	optimal value function
w_n^τ	objective function for finite-horizon $\mathrm{AV@R}_\tau$-criterion for the total discounted cost
w^τ	objective function for infinite-horizon $\mathrm{AV@R}_\tau$-criterion for the total discounted cost
T_μ	one-step operator according to the stationary policy μ
T	minimizing one-step operator
$\mathscr{R}$	strongly communicating class
$\mathscr{S}$	set of sinks
$\mathscr{T}$	set of states which are transient under every policy

Bibliography

Ravi P. Agarwal, Donal O'Regan, and D. R. Sahu. *Fixed Point Theory for Lipschitzian-type Mappings with Applications*, volume 6 of *Topological Fixed Point Theory and Its Applications*. Springer, 2009.

Eitan Altman. *Constrained Markov Decision Processes*. Stochastic Modeling. Chapman & Hall/CRC, 1999.

Eitan Altman and Shaler Stidham, Jr. Optimality of monotonic policies for two-action Markovian decision processes, with applications to control of queues with delayed information. *Queueing Systems*, 21(3-4):267–291, September 1995.

P. S. Ansell, K. D. Glazebrook, J. Niño-Mora, and M. O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57(1):21–39, April 2003.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, July 1999.

Nicole Bäuerle and André Mundt. *Risikomanagement*, volume 3 of *Schriftenreihe des Kompetenzzentrums Versicherungswissenschaften*. Verlag Versicherungswirtschaft GmbH Karlsruhe, 2005. In German.

Melike Baykal-Gürsoy and Keith W. Ross. Variability Sensitive Markov Decision Processes. *Mathematics of Operations Research*, 17(3):558–571, August 1992.

Richard Bellman. On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38:716–719, 1952.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 2nd edition, 2001.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, 3rd edition, 2005.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, 1978.

Frederick J. Beutler and Keith W. Ross. Uniformization for Semi-Markov Decision Processes under Stationary Policies. *Journal of Applied Probability*, 24(3):644–656, September 1987.

Patrick Billingsley. *Probability and Measure*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., third edition, 1995.

Kang Boda and Jerzy A. Filar. Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63(1):169–186, February 2006.

Kang Boda, Jerzy A. Filar, Yuanlie Lin, and Lieneke Spanjers. Stochastic Target Hitting Time and the Problem of Early Retirement. *IEEE Transactions on Automatic Control*, 49(3):409–419, March 2004.

M. Bouakiz and Y. Kebir. Target-Level Criterion in Markov Decision Processes. *Journal of Optimization Theory and Applications*, 86(1):1–15, July 1995.

Mokrane Bouakiz and Matthew J. Sobel. Inventory Control with an Exponential Utility Criterion. *Operations Research*, 40(3):603–608, May-June 1992.

Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

Hyeong Soo Chang, Michael C. Fu, Jiaqiao Hu, and Steven I. Marcus. *Simulation-based Algorithms for Markov Decision Processes*. Springer, 2007.

Kun-Jen Chung and Matthew J. Sobel. Discounted MDP's: Distribution Functions and Exponential Utility Maximization. *SIAM Journal Control and Optimization*, 25(1):49–62, January 1987.

Thomas B. Crabhill. Optimal Control of a Maintenance System with Variable Service Rates. *Operations Research*, 22(4): 736–745, July-August 1974.

C. Cromvik and M. Patriksson. On the Robustness of Global Optima and Stationary Solutions to Stochastic Mathematical Programs with Equilibrium Constraints, Part 2: Applications. *Journal of Optimization Theory and Applications*, 144 (3):479–500, March 2010.

D. P. de Farias and B. van Roy. The Linear Programming Approach to Approximate Dynamic Programming. *Operations Research*, 51(6):850–865, November-December 2003.

Kresimir Delac and Mislav Grgic. A Survey of Biometric Recognition Methods. In *46th International Symposium Electronics in Marine*, pages 184–193, June 2004.

S. D. Deshmukh and Wayne Winston. Stochastic Control of Competition through Prices. *Operations Research*, 27(3): 583–594, May-June 1979.

Fabrice Dusonchet and Max-Olivier Hongler. Continuous-Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production. *IEEE Transactions on Robotics and Automation*, 19(6):977–990, December 2003.

Antonis Economou. On the control of a compound immigration process through total catastrophes. *European Journal of Operational Research*, 147(3):522–529, June 2003.

J. S. Fang, M. S. Mannan, D. M. Ford, J. Logan, and A. Summers. Value at Risk Perspective on Layers of Protection Analysis. *Process Safety and Environmental Protection*, 85(1):81–87, 2007.

Eugene A. Feinberg and Adam Shwartz. Markov Decision Models with Weighted Discounted Criteria. *Mathematics of Operations Research*, 19(1):152–168, February 1994.

Youyi Feng and Zhan Pang. Dynamic Coordination of Production Planning and Sales Admission Control in the Presence of a Spot Market. *Naval Research Logistics*, 57:309–329, 2010.

Jerzy A. Filar, Dmitry Krass, and Keith W. Ross. Percentile Performance Criteria For Limiting Average Markov Decision Processes. *IEEE Transactions on Automatic Control*, 40(1):2–10, January 1995.

Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. de Gruyter, 2004.

Auroop R. Ganguly, João Gama, Olufemi A. Omitaomu, Mohamed Medhat Gaber, and Ranga Raju Vatsavai, editors. *Knowledge Discovery from Sensor Data*. Industrial Innovation Series. CRC Press, 2009.

J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–164, 1979.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient Solution Algorithms for Factored MDPs. *Journal of Artificial Intelligence Research*, 19(1):399–468, July 2003.

Xianping Guo and Onésimo Hernández-Lerma. *Continuous-Time Markov Decision Processes: Theory and Applications*. Number 62 in Stochastic Modelling and Applied Probability. Springer, 2009.

Xianping Guo, Qiuli Liu, and Tianshou Zhou. Optimal Control of Stochastic Fluctuations in Biochemical Reactions. *Journal of Biological Systems*, 17(2):283–301, June 2009.

Diwakar Gupta and Lei Wang. Revenue Management for a Primary-Care Clinic in the Presence of Patient Choice. *Operations Research*, 56(3):576–592, May-June 2008.

Onésmio Hernández-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control: Basic Optimality Criteria*. Number 30 in Applications of Mathematics. Springer, 1996.

Ronald A. Howard and James E. Matheson. Risk-Sensitive Markov Decision Processes. *Management Science*, 18(7): 356–369, March 1972.

Stratton C. Jaquette. Markov Decision Processes with a New Optimality Criterion: Discrete Time. *The Annals of Statistics*, 1(3):496–505, May 1973.

Stratton C. Jaquette. Markov Decision Processes with a New Optimality Criterion: Continuous Time. *The Annals of Statistics*, 3(2):547–553, March 1975.

Stratton C. Jaquette. A Utility Criterion for Markov Decision Processes. *Management Science*, 23(1):43–49, September 1976.

Jin Fai Kan and Christian R. Shelton. Solving Structured Continuous-Time Markov Decision Processes. In *Tenth International Symposium on Artificial Intelligence and Mathematics*, 2008.

Roger Koenker. *Quantile Regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, 2005.

E. G. Kyriakidis. On the control of a truncated general immigration process through the introduction of a predator. *Journal of Applied Mathematics and Decision Sciences*, 2006:1–12, 2006.

Claude Lefèvre. Optimal Control of a Birth and Death Epidemic Process. *Operations Research*, 29(5):971–982, September-October 1981.

Juan Liu, Chunhua Men, Victor E. Cabrera, Stan Uryasev, and Clyde W. Fraisse. Optimizing Crop Insurance under Climate Variability. *Journal of Applied Meteorolgy and Climatology*, 47:2572–2580, October 2008.

Elisa Luciano, Lorenzo Peccati, and Donato M. Cifarelli. VaR as a risk measure for multiperiod static inventory models. *International Journal of Production Economics*, 81–82(1):375–384, January 2003.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press, 2005.

Suresh, K. Nair and Ravi Bapna. An Application of Yield Management for Internet Service Providers. *Naval Research Logistics*, 48:348–362, 2001.

Evan L. Porteus. On the Optimality of Structured Policies in Countable Stage Decision Processes. *Management Science*, 22(2):148–157, October 1975.

Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2007.

L. Pruzzo, R. J. C. Cantet, and C. C. Fioretti. Risk-adjusted expected return for selection decisions. *Journal of Animal Science*, 81(12):2984–2988, December 2003.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2005.

Qinru Qiu and Massoud Pedram. Dynamic Power Management Based on Continuous-Time Markov Decision Processes. In *Design Automation Conference*, pages 555–561, 1999.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, July 2002.

Keith W. Ross and Ravi Varadarajan. Multichain Markov Decision Processes with a Sample Path Constraint: A Decomposition Approach. *Mathematics of Operations Research*, 16(1):195–207, February 1991.

Dwight R. Sanders and Mark R. Manfredo. The Role of Value-at-Risk in Purchasing: An Application to the Foodservice Industry. *The Journal of Supply Chain Management*, 38(2):38–45, May 2002.

Manfred Schäl. *Markoffsche Entscheidungsprozesse*. Teubner-Skripten zur mathematischen Stochastik. B. G. Teubner, 1990. In German.

Manfred Schäl. On Discrete-Time Dynamic Programming in Insurance: Exponential Utility and Minimizing the Ruin Probability. *Scandinavian Actuarial Journal*, 3:189–210, 2004.

Richard F. Serfozo. An Equivalence Between Continuous and Discrete Time Markov Decision Processes. *Operations Research*, 27(3):616–620, May-June 1979.

Chiao-Fe Shu, Arun Hampapur, Max Lu, Lisa Brown, Jonathan Connell, Andrew Senior, and Yingli Tian. IBM Smart Surveillance System (S3): A Open and Extensible Framework for Event Based Surveillance. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, pages 318–323, September 2005.

Peter Singer. *Animal Liberation*. HarperCollinsPublishers, 2002.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education, Inc., 2006.

Charles S. Tapiero. Value at risk and inventory control. *European Journal of Operational Research*, 163(3):769–775, June 2003.

Massimo Tistarelli and Mark S. Nixon, editors. *Advances in Biometrics: Third International Conference, ICB 2009, Proceedings*. Springer, 2009.

Massimo Tistarelli, Stan Z. Li, and Rama Chellappa, editors. *Handbook for Remote Biometrics for Surveillance and Security*. Advances in Pattern Recognition. Springer, 2009.

William F. Trench. *Introduction to Real Analysis*. October 2009. Free Edition 1.01, available under `http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_REAL_ANALYSIS.PDF`.

Peter M. Verderame and Christodoulos A. Floudas. Operational Planning of Large-Scale Industrial Batch Plants under Demand Due Date and Amount Uncertainty: II. Conditional Value-at-Risk Framework. *Industrial & Engineering Chemical Research*, 49(1):260–275, January 2010.

R. B. Webby, P. T. Adamson, J. Boland, P. G. Howlett, A. V. Metcalfe, and J. Piantadosi. The Mekong—applications of value at risk (VaR) and conditional value at risk (CVaR) simulation to the benefits, costs and consequences of water resources development in a large river basin. *Ecological Modelling*, 201(1):89–96, February 2007.

D. J. White. Mean, Variance, and Probabilistic Criteria in Finite Markov Decision Processes: A Review. *Journal of Optimization Theory and Applications*, 56(1):1–29, January 1988.

D. J. White. A Survey of Applications of Markov Decision Processes. *The Journal of the Operational Research Society*, 44(11):1073–1096, November 1993.

P. Whittle. Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, 25:287–298, 1988.

Congbin Wu and Yuanlie Lin. Minimizing Risk Models in Markov Decision Processes with Policies Depending on Target Values. *Journal of Mathematical Analysis and Applications*, 231(1):47–67, March 1999.