Monotonicity results for multi-armed bandit processes with multiple active actions

Peter Sieb

December 7, 2012[*]

Abstract

To address the allocation of scarce resources on various projects which are driven by partially observable processes, we develop a 2-step optimization scheme. Based on the uncertain project stages and the waiting times for activation, it must be decided which projects should be operated actively and, hereupon, which operation mode - wait, test or intervene - should be selected. This problem can be decomposed in an operational and an allocation decision which both can be optimized by means of an infinite-horizon partially observable Markov decision process. As in general the latter is computationally intractable, we present a heuristic allocation rule based on the well-known MAB approach. Assuming stable projects, we establish structural results for both the optimal operational and the heuristic allocation decision rule. In particular, the optimal operational plan is characterized by an at most 3-action region rule which can be transferred to the allocation rule under additional assumptions.

Keywords: Partially observable Markov decision processes; Dynamic programming; Multi-armed bandit processes; Scheduling

1 Introduction

Multi-armed bandit (MAB) processes refer to a class of dynamic decision-making processes, dealing with the problem of allocating scarce resources to a set of independent, competing projects (often called arms or bandits) which can be described by controllable stochastic processes. At any decision period, the decision-maker selects for each process a specific action, which consumes a certain amount of resources and controls the stochastic evolution as well as the performance of the process. In general, there are two actions available that are classified either as active or passive depending on their resource consumption and consequences. The objective of the decision-maker is to allocate the resources in such a way that the system's

[*]Institut für Operations Research, Universität Karlsruhe, D-76128 Karlsruhe, Germany

performance is maximized. As such a problem, in general, is intractable, the solution of MAB processes resorts to so-called index rules: Firstly, for each project a state-dependent index is determined and, afterwards, the resources are allocated in the order of these index values.

In this paper, we present a general framework for allocating resources to a set of non-identical projects which can be operated in different modes. Hereby, we draw on a partially observable version of the general MAB processes introduced by Glazebrook and Minty (2009) and extend this framework to multiple active actions, i.e. for each project an intervention and a test is available. As an example, consider a machine maintenance problem where each machine can be either inspected or repaired, or a patient scheduling problem where each patient can be either tested or operated. As Glazebrook (1982) proves within the classic MAB framework of Gittins and Jones (1979), this extension of the action space does not influence the quality of the heuristic if certain assumptions are satisfied. Another feature of our model is that we do not just focus on the project's partially observable stage, but also on the time which has passed since the last active action has been scheduled.

Our general problem formulation can be decomposed in an operational level, where an optimal action is assigned to each project's stage, and an allocation level, where resources are assigned to individual projects which on the other hand are necessary to execute the optimal operational action. In particular, we first of all formulate a partially observable Markov decision process (POMDP) for each project in order to decide whether no action, a test or an intervention should be executed. Afterwards, on the allocation level we decide by means of a partially observable MAB process which project should be operated according to its optimal operational action and which project should be operated passively instead. The objective of both optimization steps is to maximize the total expected reward. Consequently, our formulation additionally allows the analysis of the mutual influence between the operational decision and the resource allocation.

The aim of this article, however, is not to provide a decision support tool that can readily be used in real time. We rather focus on capturing the most essential structures of the decision problem. Specifically, we establish structural results for both the optimal operational decision rule and the heuristic allocation decision rule. In particular, we show that under the assumption of stable projects the operational decision is characterized by an at maximum 3-action region rule. Furthermore, we are able to identify quite general conditions which ensure that the index-based allocation decision rule follows a monotone switching curve rule. Combining these results, we finally obtain a highly structured at most 3-action region allocation rule which is monotone in the individual project's stage and its waiting time.

Outline The remainder of this article is organized as follows: Section 2 gives account of previous work. In section 3, the underlying decision problem is presented rigorously. In section 4, we formulate the optimization problem by decomposing it into an operational and an allocation level. For the latter, we prove that the allocation problem falls within the class of MAB processes. Our main result, the optimality of an at most 3-action region rule for both the operational and the allocation decision problem, is then proven in section 5.

Notation. We use $\mathbb{R}$ ($\mathbb{R}_+$) to denote the set of (nonnegative) real numbers and $\mathbb{N}_0$ ($\mathbb{N}$) to denote the set of nonnegative (positive) integers. Considering $(W^1, \leq_1)$ and $(W^2, \leq_2)$, a real-valued function $v : W^1 \to W^2$ is said to be increasing (decreasing) on $(W^1, \leq_1)$ if $x \leq_1 x'$ implies $v(x) \leq_2 v(x')$ ($v(x) \geq_2 v(x')$). By e_n^N, we provide an element of $\mathbb{R}^N$ with 1 at the n-th position and 0 at all other positions (nth canonical unit vector of length N).

2 Related Literature

As our model is built on the theory of MAB processes and, additionally, bears some resemblance to the machine maintenance literature, our literature review refers to both fields.

Multi-armed bandit processes

A comprehensive introduction to MAB processes can be found in Mahajan and Teneketzis (2008) or Gittins et al. (2011). In the classic version of the MAB framework, the resource capacity as well as the resource consumption of an active action is restricted to one. If a process is not activated, no resources are consumed and no reward is incurred. As a result of an allocation decision, only the state of the process that is activated changes. Gittins and Jones (1979) showed for this classic allocation problem that an index can be assigned to each process as a function of the current state and that an optimal decision rule activates the process with the largest index. Whittle (1988) proposed an extension of the classic MAB model, the so-called restless MAB processes, by incorporating an arbitrary resource capacity. In addition, passive projects may also yield a non-zero reward and change their states. Similar to the classic approach, Whittle (1988) derived a state-dependent index for each process and suggested that the available resources should be allocated in the order of the index values. This index-based decision rule provides a powerful heuristic which is known to be asymptotically optimal under certain assumptions (see Weber and Weiss, 1990). Whittle's model and its extensions have been proposed in many application contexts. These include the routing of unmanned military aircrafts (see, for example, Le Ny et al., 2008), inventory routing (Archibald et al., 2009), machine maintenance (e.g. Glazebrook et al., 2005) and queuing control (Ansell et al., 2003). Finally, Glazebrook and Minty (2009) generalized the restless bandit processes of Whittle (1988) by additionally incorporating an arbitrary resource consumption. The heuristic developed by Whittle (1988) can readily be transferred to this extended framework. Finally, in the context of partially observable Markov decision processes, we want to highlight the work of Krishnamurthy and Wahlberg (2009) who transferred the classic MAB framework of Gittins and Jones (1979) to POMDPs and showed by means of the likelihood ratio order that under certain assumptions the resulting indices are monotone in the projects' states.

Machine maintenance in a partially observable environment

The classic machine maintenance problem deals with the optimal maintenance of a machine whose partially observable condition gradually deteriorates. In addition to a repair action that improves the machine's condition, an inspection action is available which delivers additional information about the actual wear state of the machine. The goal is to determine a decision rule that minimizes the cost of the overall machine maintenance. An example of such a model can be found in Monahan (1980). In this work, Monahan (1980) proved that the optimal decision rule for this problem is in general unstructured. For the special case of complete information, several authors, however, showed by means of the likelihood ratio order that an optimal decision rule is composed of at most four action regions (e.g. Ohnishi et al., 1986; Jin et al., 2005). A detailed overview of various structural results for maintenance models can be found in Zheltova (2010).

3 The model

In an infinite-horizon decision problem, in each decision period C resources are available to operate M heterogeneous projects, where in general the total resource consumption exceeds the capacity if each project is operated actively. Each project $b \in B := \{1, \ldots, M\}$ is driven by a partially observable process. In general, we label the process parameters of an individual project by the superscript b, but since the following discussion deals only with a single project, we will drop this superscript in this section.

Partially observable stage process

An individual project evolves through stages $i \in I = \{1, \ldots, N\}$, where N denotes an absorbing stage in which no further activation is appropriate or possible (e.g. a machine breaks down or a patient dies). Thus, stage N represents an absorbing stage in which the operational process terminates. The current stage of the project, however, cannot be directly observed. Instead the decision-maker makes a stage observation $\theta \in I$ which allows him to estimate the actual stage.

For each project, there are three operation modes - wait, test and intervene - available, which are summarized in the action space $A = \{a_W, a_T, a_I\}$. The execution of each action $a \in A$ is associated with a resource consumption $c(a)$. Specifically, we assume $c(a_T), c(a_I) > c(a_W) = 0$ which reflects the fact that both a_T and a_I represent an active engagement in the project's development whereas a_W is a passive action. As a further consequence of the execution of an action a in project stage i, the decision-maker collects a reward $r(i, a)$. As action a_W does neither have an impact on the project's development nor delivers additional information about the project's stage, we assume that no reward incurs, i.e. $r(\cdot, a_W) = 0$. The execution of the test a_T basically has the purpose of acquiring more precise information about the current stage of the project. This information gain is associated with a stage-independent reward $r(\cdot, a_T) = const.$. Whether $r(\cdot, a_T)$ is positive or negative depends upon the specific model application. While in general $r(\cdot, a_T)$ can be treated as costs, for example, in a health care context performing a test might earn some money for the health care provider. Lastly, a_I represents a final intervention such as, for example, the replacement of a machine or a surgery of a patient, which terminates the project's development. We hereby assume that the benefit of the intervention is the greater, the more the project has progressed, i.e. the higher the index of its stage is. Thus, we have $r(N-1, a_I) \geq \ldots \geq r(1, a_I)$. As stage N represents the project's end, we finally assume $r(N, a) = 0$ for any $a \in A$.

After having collected the one-stage reward, the project transitions to stage j with probability $p^a(i, j)$, where the decision-maker obtains an observation θ with probability $q^a(j, \theta)$. We presume that, if the project's development is not finalized by a_I, it terminates and transitions into stage N with stage-independent probability ϖ. In this case, the collection of the terminal reward is not possible anymore. If, however, the project is operated by a_I, it indeed transitions into stage N, but, before, the decision-maker is able to realize the reward $r(i, a_I)$. Finally, we assume that, if the project has reached its terminal stage, the decision-maker receives an observation which provides complete information. Altogether, it then holds that $p^a(i, N) = \varpi$ for any $a \in \{a_W, a_T\}, i \neq N$, as well as $p^{a_I}(i, N) = 1$, $p^a(N, N) = 1$, and $q^a(N, N) = 1$ for any $a \in A$.

Belief state representation

It is well known that a partially observable Markov decision process can be reduced to a fully observable decision process (see, for example, Sondik, 1978) by means of a belief state: As the project's stage is only partially known, the decision-maker needs to rely on an estimate which can be expressed by a belief state $x = (x_1, \ldots, x_N) \in X(I)$, where $X(I)$ denotes the set of probability distributions over I. Such a belief state assigns a probability x_i to each stage $i \in I$. Then, if action a is executed in belief state x, a reward $\hat{r}(x, a) = \sum_i x_i r(i, a)$ is realized. Analogously, in belief state x observation θ is made with probability $\rho^a(x, \theta) = \sum_{i,j} x_i p^a(i, j) q^a(j, \theta)$. As a consequence of the resulting observation, the project transitions into belief state $y = T(x, a, \theta)$, where

$$T(x, a, \theta)_j = \frac{\sum_i x_i p^a(i, j) r^a(j, \theta)}{\rho^a(x, \theta)}, j \in I.$$

Waiting time and penalty costs

We assume that for each project an optimal operational plan $f^{OP*} : X(I) \to A$, which assigns a treatment to each belief state, is given. This operational plan serves as a guideline for the allocation planning process. Moreover, a project is characterized by its current waiting time for activation, which we denote by $w \in \mathbb{N}_0$. The current waiting time w causes waiting costs $k(w) \geq 0$, which we assume to be increasing in w and bounded from above.

This concept reflects the fact that an increasing idle time of a project causes additional costs as, for example, customers waiting for service grow dissatisfied. The waiting time of the project evolves as follows: If a_W is selected, although an active action is optimal, the waiting time increases by 1, otherwise, it is reset to 0. Accordingly, the waiting time is set to $\zeta((x,w),a_W) = w+1$ in the case of $f^{OP*}(x) \neq a_W$, and $\zeta((x,w),a) = 0$ otherwise. Finally, penalty costs $\kappa(x,a) >> 0$ incur if an active action $a \in \{a_T, a_I\}$ is selected, although the optimal operational plan proposes to execute $f^{OP*}(x) \neq a$. As f^{OP*} is supposed to serve as a strict guideline, we assume that the penalty costs are high enough to ensure that it is never optimal to select an active action $a \neq f^{OP*}(x)$.

4 Optimization

Before we turn to the capacity allocation problem, we first of all point out how to determine an optimal operational plan for an individual project.

4.1 Operational planning

As we only focus on the operational decision process for an individual project in this section, we forego the superscript b again. In order to determine an optimal operational plan, we formulate a belief state MDP resulting from the belief state reduction of the original partially observable decision process. This MDP consists of state space $X(I)$, action set A, transition probability $\rho^a(x,\theta)$ from belief state x to $T(x,a,\theta)$, one-stage reward $\hat{r}(x,a)$ and discount factor β.

A stationary (Markov) policy $\pi^{OP} = (f^{OP}, f^{OP}, \ldots)$ is defined as a sequence of identical decision rules $f^{OP} : X(I) \to A$ specifying the action $a = f^{OP}(x)$ to be taken. As, according to Blackwell (1965), for each belief state MDP there exists an optimal stationary policy, it is sufficient to determine an optimal decision rule f^{OP*}. Let F^{OP} be the set of all operational decision rules. Denote by $(\mathfrak{X}_t, t \in \mathbb{N}_0)$ the state process of the belief state MDP and introduce $V^{OP}(x)$ to be the maximal expected total reward starting in belief state x. Then $V^{OP}(x)$ is defined by

$$V^{OP}(x) = \max_{f^{OP} \in F^{OP}} \mathbb{E}_{f^{OP}} \left[\sum_{t=0}^{\infty} \beta^t \hat{r}(\mathfrak{X}_t, f^{OP}(\mathfrak{X}_t)) | \mathfrak{X}_0 = x \right], x \in X(I).$$

It is well known in dynamic programming that V^{OP} is the unique solution to the optimality equation

$$V^{OP}(x) = \max_{a \in A} \left\{ L^{OP} V^{OP}(x,a) \right\}, x \in X(I), \tag{1}$$

where

$$L^{OP} V^{OP}(x,a) := \hat{r}(x,a) + \beta \sum_{\theta \in I} \rho^a(x,\theta) V^{OP}(T(x,a,\theta)), a \in A.$$

Moreover, each decision rule f^{OP*} formed by actions $f^{OP*}(x)$ maximizing the right hand side of (1) is optimal, i.e. leads to V^{OP}.

Since stage N is absorbing and we assume $r(N, a) = 0$, without loss of generality we can set $f^{OP*}(e_N^N) := a_W$. Furthermore, we hereafter assume that the project starts in a belief state $x \in X(I)$ where $x_N = 0$. Due to $q^a(N, N) = 1$ for any $a \in A$, the probability for stage N, i.e. x_N, is either 0 or 1. Consequently, we can focus our analysis on the essential belief state space $\bar{X}(I) := \{x \in X(I) | x_N = 0\}$.

4.2 Capacity allocation

Readopting our superscript notation, the allocation decision process can be summarized as follows:

i) The state of the system is a vector $(\mathbf{x}, \mathbf{w}) = ((x^1, w^1), \ldots, (x^M, w^M)) \in \mathcal{X}$, where $\mathcal{X} := \times_{b \in B}(X(I^b) \times \mathbb{N}_0)$.

ii) The set of allocation actions is a M-vector $\mathbf{a} \in A^M$, where $\mathbf{a}$ is a tuple $\mathbf{a} := (a^1, \ldots, a^M)$ which comprises an action for each project.

iii) As a consequence of executing action $\mathbf{a}$, a reward is realized as follows:

$$\hat{\mathbf{r}}((\mathbf{x}, \mathbf{w}), \mathbf{a}) := \sum_{b \in B} \hat{r}^b(x^b, a^b) - k^b(w^b) - \kappa^b(x^b, a^b).$$

Note that rewards are bounded and $\beta \in (0, 1)$ is the discount factor.

iv) As a further consequence of action $\mathbf{a}$, an observation $\theta^b \in I^b$ is made with probability $\rho^{a^b,b}(x^b, \theta^b)$ for each $b \in B$. Subsequently, the state of each project evolves independently of each other to $(T^b(x^b, a^b, \theta^b), \zeta^b((x^b, w^b), a^b)$, where

$$\zeta^b((x^b, w^b), a^b) := \begin{cases} w^b + 1 & , f^{OP,b*}(x^b) \in \{a_T, a_I\} \wedge a^b = a_W, \\ 0 & , \textit{otherwise.} \end{cases}$$

v) There are C resources available which leads to the resource constraint $\sum_{b \in B} c^b(a^b) \leq C$.

The resource allocation problem can be modeled by means of a constrained belief state MDP. Denote by $\mathfrak{X}_t = (\mathfrak{X}_t^1, \ldots, \mathfrak{X}_t^M)$ and $\mathfrak{w}_t = (\mathfrak{w}_t^1, \ldots, \mathfrak{w}_t^M), t \in \mathbb{N}_0$, random variables with realizations $\mathbf{x}_t \in \mathcal{X}$ and $\mathbf{w}_t \in \mathbb{N}_0^M$ which indicate the belief state and the waiting time of each project at time t, respectively. Let $f^{AP} : \mathcal{X} \to A^M$ be a decision rule for the allocation problem, and F^{AP} be the set of all allocation rules. Then, starting in state $(\mathbf{x}, \mathbf{w})$ the maximal expected total reward is given by

$$\sup_{f^{AP} \in F^{AP}} \mathbb{E}_{f^{AP}} \left[\sum_{t=0}^{\infty} \beta^t \hat{\mathbf{r}}((\mathfrak{X}_t, \mathfrak{w}_t), f(\mathfrak{X}_t, \mathfrak{w}_t)) \big| (\mathfrak{X}_t, \mathfrak{w}_t) = (\mathbf{x}, \mathbf{w}) \right]$$
$$s.t. \sum_{b \in B} c^b(f(\mathbf{x}_t, \mathbf{w}_t)) \leq C, (\mathbf{x}_t, \mathbf{w}_t) \in \mathcal{X}, t \in \mathbb{N}_0.$$

The state space of this problem formulation, however, is growing exponentially in the number of projects. To that effect, which is also known as *curse of dimensionality*, Papadimitriou and Tsitsiklis (1999) suggest that even the solution of a deterministic version of such an allocation problem is PSPACE-hard. Correspondingly, a pure dynamic programming approach is unlikely to be insightful and may be computationally intractable for problems of reasonable size. Accordingly, our primary task is to develop a good heuristic decision rule. In fact, the constrained MDP presented above falls within the class of multi-armed bandit problems, a famously intractable class of scheduling models introduced by Gittins and Jones (1979).

4.2.1 Heuristic allocation rule

In one of the latest extensions of the MAB class, Glazebrook and Minty (2009) present a similar model to ours and develop project-specific priority indices, $J^b : X(I^b) \times \mathbb{N}_0 \to \mathbb{R}$. The resulting index-based allocation rule allocates the resources as follows: The projects are ordered according to the value of their priority indices $J^b(x^b, w^b)$ and, beginning with the first, the projects are activated by executing the optimal action $f^{OP,b*}(x^b)$, as long as resources are left and their priority indices are positive.

Allocation indices

In this section, we derive the priority index J^b for an individual project and, therefore, drop the superscript b again. Firstly, we introduce project-specific resource costs ν for each consumed resource unit. Thus, in addition to the given reward structure there incur resource costs $\nu c(a)$. The priority index J^b is then defined as the break-even resource costs which establish indifference between the optimal action $f^{OP*}(x)$ and the passive action a_W. In order to derive this crucial result, we firstly model an individual project process by means of a MDP as follows:

i) The state of the project is a tuple $(x, w, \nu) \in X(I) \times \mathbb{N}_0 \times \mathbb{R}$.

ii) The action set is A.

iii) As a consequence of performing action a, a reward is incurred as follows:

$$\tilde{r}((x, w, \nu), a) := \hat{r}(x, a) - k(w) - \kappa((x, \nu), a) - \nu c(a).$$

Note that rewards are bounded, $\beta \in (0, 1)$ is the discount factor and $\kappa((x, \nu), a)$ is assumed to be high enough to ensure that an action $a \in \{a_T, a_I\}$ is suboptimal if $a \neq f^{OP}(x)$.

iv) The decision-maker obtains an observation $\theta \in I$ with probability $\rho^a(x, \theta)$ and the state of the project evolves to $(T(x, a, \theta), \zeta((x, w), a), \nu)$.

Let $f : X(I) \times \mathbb{N}_0 \times \mathbb{R}_0 \to A$ be a decision rule, and F be the set of all decision rules. Then, the maximal expected total reward starting in state (x, w, ν) is defined by

$$V(x, w, \nu) = \max_{a \in A} \{LV((x, w, \nu), a)\}, (x, w, \nu) \in X(I) \times \mathbb{N}_0 \times \mathbb{R}_0, \tag{2}$$

where

$$LV((x, w, \nu), a) := \tilde{r}((x, w, \nu), a) + \beta \sum_{\theta \in I} \rho^a(x, \theta) V(T(x, a, \theta), \zeta((x, w), a), \nu).$$

For each state (x, w, ν) there exists a dominant active action given by

$$a^*_{act}(x, w, \nu) := \arg \max_{a \in \{a_T, a_I\}} \{LV((x, w, \nu), a)\}.$$

According to Glazebrook (1982), we need to make sure that $a^*_{act}(x, w, \nu)$ is independent of ν in order to establish an index-based decision rule. Consider state (x, w, ν) and assume $f^{OP*}(x) = a_W$. Because of the penalty costs and assumption (A1), for the considered project it is always optimal to execute action a_W in the current and any future decision period. Therefore, it is sufficient to concentrate on the essential state space $\mathfrak{P} \times \mathbb{R}_0$, where $\mathfrak{P} := \{(x, w) \in \bar{X}(I) \times \mathbb{N}_0 | f^{OP*}(x) \in \{a_T, a_I\}\}$. If we restrict our analysis to $\mathfrak{P}$, the penalty costs ensure that $a^*_{act}(x, w, \nu) = f^{OP*}(x)$ and, thus, $a^*_{act}(x, w, \nu)$ is independent of ν. Furthermore, the following property, which is well-known as indexability, must be fulfilled.

Definition 4.1. *(Whittle, 1988)* A project is said to be *indexable* if the set of states where the passive action is optimal in the single-project subproblem (2) increases monotonically from the empty set to the full set of states as ν increases from $-\infty$ to ∞.

In order to meet this requirement, we focus on stable projects by the following assumption.

(A1) It holds that $T(x, a_W, \theta) = x$ for any $x \in X(I), \theta \in I \backslash \{N\}$ as well as $w^{a_W}(x, N) = \varpi$ for any $x \in \bar{X}(I)$.

If it is optimal to select action a_W in belief state x and (A1) is valid, the belief state can only change by transitioning into the absorbing belief state e^N_N. Note that by (A1) we do not assume that the actual project stage remains constant, but that the belief state does not change. This assumption, for example, is satisfied if the underlying stage does not undergo a definite development and the observations resulting from a_W do not allow a conclusion about the actual project stage. In the remainder of this paper, we only consider stable projects.

Lemma 4.1. *Assume (A1). Then, it holds that*

$$\frac{\partial LV((x, w, \nu), f^{OP*}(x))}{\partial \nu} \leq \frac{\partial LV((x, w, \nu), a_W)}{\partial \nu}.$$

Proof. If in state (x, w, ν) action a_W is selected and (A1) is valid, the belief state remains the same, unless it transitions into the absorbing state e_N^N. As we assume $f^{OP*}(x) \in \{a_T, a_I\}$, it is optimal to either select a_W in each future period or to select $f^{OP*}(x)$ after $\tau \in \mathbb{N}_0$ periods, unless the project has transitioned into the absorbing state. Altogether, we can conclude that after selecting a_W in state (x, w, ν) action $f^{OP*}(x) = a$ will be selected after $\tau \in \mathbb{N}_0 \cup \infty$ additional periods. This yields

$$\begin{aligned}\frac{\partial LV((x,w,\nu),a_W)}{\partial \nu} &= \sum_{t=0}^{\tau} \beta^t c(a_W) + (\beta(1-\varpi))^{\tau+1} \frac{\partial LV((x,w+\tau+1,\nu),a)}{\partial \nu} \\ &\geq \frac{\partial LV((x,w+\tau+1,\nu),a)}{\partial \nu} \\ &= \frac{\partial LV((x,w,\nu),a)}{\partial \nu}.\end{aligned}$$

□

Lemma 4.2. *Assume (A1). Then, the project's process is indexable.*

Proof. For any (x, w) it holds that $\nu \to -\infty \Rightarrow f^*(x, w, \nu) = f^{OP*}(x) \in \{a_T, a_I\}$ and $\nu \to \infty \Rightarrow f^*(x, w, \nu) = a_W$. Applying Lemma 4.1, the indexability property is fulfilled. □

Having proven indexability, we now define the priority index J as the break-even resource costs given by

$$J(x, w) := \inf \left\{ \nu \in \mathbb{R}_0 | LV((x, w, \nu), f^{OP*}(x)) = LV((x, w, \nu), a_W) \right\}.$$

The priority index represents the additional reward achieved by the dominant active action in relation to the additional resource consumption.

Index-based allocation rule

Back on the allocation level, Glazebrook and Minty (2009) suggest to carry out the dominant active action of those projects with the largest index values in descending order, until either the resource capacity is used up or only processes with non-positive index values remain. For the rest of the projects the passive action is selected. Subsequently, a project is operated actively if its priority index is high enough, i.e. exceeds a particular threshold which we refer to as critical index.

Readopting our superscript notation, this critical index of project b, which we denote by $\Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b})$, is dependent of the states of the other projects $(\mathbf{x}^{-b}, \mathbf{w}^{-b})$, where $\mathbf{x}^{-b} := (x^1, \ldots, x^{b-1}, x^{b+1}, \ldots, x^M)$ and $\mathbf{w}^{-b}$ analogously. For any state $(\mathbf{x}, \mathbf{w}) \in (\times_{b \in B} \mathfrak{P}^b)$ the index-based allocation rule is then given by $f^I = (f^{I,1}, \ldots, f^{I,M}) \in F$, where

$$f^{I,b}(\mathbf{x}, \mathbf{w}) := \begin{cases} a_W & , J^b(x^b, w^b) < \Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b}), \\ f^{OP,b*}(x^b) & , J^b(x^b, w^b) \geq \Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b}). \end{cases}$$

The proposed index rule is characterized by a boundary function below which a_W and above which $f^{OP,b*}(x^b)$ is selected. Such a decision rule is commonly referred to as *switching curve decision rule* (Lewis, 2001).

Note that, since the index of a project is independent of the other projects, we can easily incorporate arrivals of additional projects into our model (cf. Whittle, 1981).

5 Structural results

In this section, we first of all derive structural results for an optimal operational plan and, therefore, initially drop the superscript notation. Consequently, we make use of the obtained results to establish structural results for the index-based allocation rule.

In order to establish structural results, we resort to the likelihood ratio order $\leq_{lr}$ which is defined as follows: For $x, y \in X(I)$ it holds that $y \leq_{lr} x$ if $x_i y_j \leq x_j y_i$ for any $i \geq j$. Furthermore, $y \leq_{lr} x$ implies $x \leq_{st} y$, where $\leq_{st}$ denotes the common stochastic order (Whitt, 1979). Throughout the following sections we additionally assume that for each $a \in A$ it holds that $p^a(i,j)p^a(i',j') \geq p^a(i',j)p^a(i,j')$ for $i, i', j, j' \in I$ where $i \geq i'$ and $j \geq j'$, as well as $q^a(i,\theta)q^a(i',\theta') \geq q^a(i',\theta)q^a(i,\theta')$ for any $i, i', \theta, \theta' \in I$ where $i \geq i'$ and $\theta \geq \theta'$. This implies that it is the more likely to transition into a stage with a high index, the higher the current stage index is, and that it is the more likely to obtain an observation, which indicates a stage with a high index, the larger the current stage index becomes.

5.1 Optimality of an at most 3-action region operational plan

It is well known from the machine maintenance literature that an optimal operational plan has the following property (see, for example, Ohnishi et al., 1986):

Lemma 5.1. *Let $x, y \in \bar{X}(I)$ be such that $x \geq_{lr} y$ and let a_I be an optimal action in belief state y. Then, there exists an optimal operational plan $f^{OP*} \in F^{OP}$ such that $f^{OP*}(y) = f^{OP*}(x) = a_I$.*

To establish a more rigorous structure for f^{OP*}, we examine the value function V^{OP} for monotonicity properties.

Lemma 5.2. *Assume (A1). Then, $V^{OP}(\cdot)$ is increasing on $(\bar{X}(I), \leq_{lr})$.*

Proof. The result easily follows by induction using the arguments of Proposition 1 in Lovejoy (1987) and the fact that the probability for transitioning into the absorbing state e_N^N is stage-independent. □

We are now able to prove the key result of this section, i.e. that there exists an optimal operational plan which divides the belief state space into at most three coherent action regions.

Proposition 5.1. *Assume (A1). Then, there exists an optimal operational plan $f^{OP*} \in F^{OP}$, which divides each segment $[y, z]$ into at most three segments, where $y, z \in \bar{X}(I)$ are such that $z \geq_{lr} y$ and the segment $[y, z]$ is defined as*

$$[y, z] := \{x \in \bar{X}(I_1) | x = \lambda y + (1 - \lambda)z, \lambda \in [0, 1]\}.$$

Under the assumption of $f^{OP*}(z) = a_I$, for any $x \in [y, z]$ there exist $\mu_1, \mu_2 \in [y, z]$ ($y \leq_{lr} \mu_1 \leq_{lr} \mu_2 \leq_{lr} z$) such that

$$f^{OP*}(x) = \begin{cases} a_W & , y \leq_{lr} x \leq_{lr} \mu_1, \\ a_T & , \mu_1 \leq_{lr} x \leq_{lr} \mu_2, \\ a_I & , \mu_2 \leq_{lr} x \leq_{lr} z. \end{cases}$$

Proof. Let $x, \tilde{x} \in \bar{X}(I)$ be such that $x \geq_{lr} \tilde{x}$. Assume there exists an optimal operational plan $f^{OP*} \in F^{OP}$ such that $f^{OP*}(x) = a_T$ and $f^{OP*}(\tilde{x}) = a_W$. Obviously, it holds that $L^{OP}V^{OP}(\tilde{x}, a_W) \geq 0$. Using Lemma 5.2, we obtain

$$\begin{aligned} V^{OP}(x) &\geq V^{OP}(\tilde{x}) \\ &= L^{OP}V^{OP}(\tilde{x}, a_T) \\ &\geq L^{OP}V^{OP}(\tilde{x}, a_T) - L^{OP}V^{OP}(\tilde{x}, a_W) \\ &\geq 0. \end{aligned}$$

After executing a_W in belief state x, the belief state either remains the same or transitions into the absorbing state e_N^N. Since it is optimal to select a_W in both of these belief states, starting in x, it is optimal to select a_W in the current and each future period. Due to $\hat{r}(x, a_W) = \hat{r}(e_N^N, a_W) = 0$, we immediately get

$$\begin{aligned} V^{OP}(x) &= L^{OP}V^{OP}(x, a_W) \\ &= \sum_{\tau=0}^{\infty} (\beta(1 - \varpi))^\tau \hat{r}(x, a_W) + \beta^\tau (1 - (1 - \varpi)^\tau) \hat{r}(e_N^N, a_W) \\ &= 0. \end{aligned}$$

We finally obtain $L^{OP}V^{OP}(\tilde{x}, a_T) - L^{OP}V^{OP}(\tilde{x}, a_W) = 0$ and, consequently, it is optimal to select action a_W in belief state $\tilde{x}$, too. Lemma 5.1 completes the proof of the Theorem. □

To give a more intuitive idea of the proposed structure, in Figure 1 we schematically map an optimal operational plan for $N = 4$ by means of a two-dimensional graph setting $x_3 = 1 - x_1 - x_2$. Using the example of the belief states y, z and $\tilde{y}, z$, it can be observed that any interval between two belief states, which are ordered with respect to $\leq_{lr}$, is divided into at most three coherent action segments. The essential belief state space can be characterized as follows: If there is a high probability that the project is in a mild stage - here stages 1 and 2 - , action a_W is optimal. If the probability for a mild stage is low, the intervention should be carried out. If we only have a vague idea of the actual stage, it is best to perform the test. Altogether, the presented decision rule meets our goal of a simple, intuitive, and realistic operational plan to a large extent.

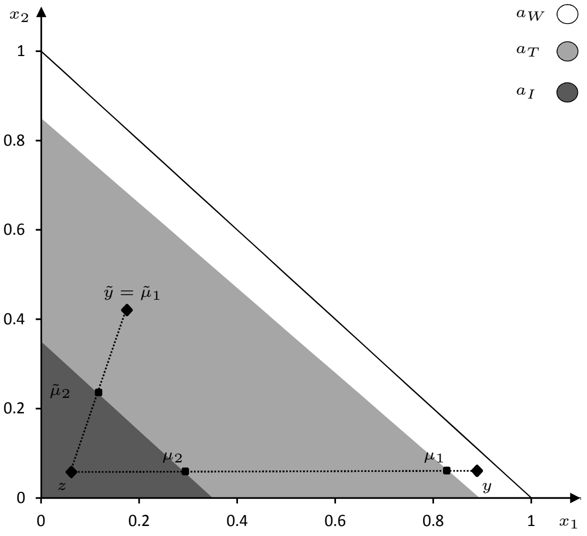

Figure 1: Schematic illustration of the decision rule presented in Proposition 5.1.

5.2 Monotone at most 3-action region allocation plan

We now analyze the structure of the index-based allocation rule. But before we turn to the allocation level, we derive monotonicity results for the priority index of individual projects. For this purpose, we initially forgo the superscript notation.

In our first step, we analyze the properties of the value function V.

Lemma 5.3. *Let $w, w' \in \mathbb{N}_0$ be such that $w \geq w'$. Then, it holds that $V(x, w, \nu) \leq V(x, w', \nu)$ for any $x \in X(I), \nu \in \mathbb{R}$.*

Proof. The proof easily follows by induction. □

Due to the penalty costs, the shape of the value function V is considerably influenced by the underlying optimal operational plan f^{OP*}. As a consequence, assuming $\nu = 0$ we are able to show that the value function V is a monotone transformation of V^{OP}.

Lemma 5.4. *Let $f^* \in F$ be an optimal decision rule. Then, there exists an optimal operational plan $f^{OP*} \in F^{OP}$ such that $f^*(x, w, 0) = f^{OP*}(x)$ for any $(x, w) \in X(I) \times \mathbb{N}_0$. Furthermore, it holds that*

$$V^{OP}(x) - k(w) = V(x, w, 0).$$

Proof. Due to the waiting costs and the penalty costs, it obviously holds that

$$V(x,w,0) \leq V^{OP}(x) - k(w).$$

Consider now an optimal operational plan $f^{OP*} \in F^{OP}$ and a decision rule $f \in F$, where $f(x,w,\nu) = f^{OP*}(x)$ for any $x \in X(I), w \in \mathbb{N}_0, \nu \in \mathbb{R}_0$. If decision rule f is executed, the waiting time will be 0 in each future period and the penalty costs will be 0 in the current as well as in each future period. This leads to the equation

$$V_f(x,w,0) = V^{OP}(x) - k(w),$$

where V_f denotes the expected total reward when implementing decision rule f. Finally, we obtain

$$V(x,w,0) = V^{OP}(x) - k(w).$$

Consequently, there exists an optimal operational plan f^{OP*} such that $f^*(x,w,0) = f^{OP*}(x)$. □

According to Lemma 5.4, there exists an optimal decision rule $f^* \in F$, which selects an active action $f^*(x,w,0) \in \{a_T, a_I\}$ in any state (x,w) that is characterized by $f^{OP*}(x) \in \{a_T, a_I\}$. Using Lemma 4.1, it follows that $J(x,w) \geq 0$. Thus, we can focus our analysis on non-negative resource costs $\nu \in \mathbb{R}_0^+$. Furthermore, using Lemma 5.2, Lemma 5.4 implies that $V(\cdot,w,0)$ is increasing on $(\bar{X}(I), \leq_{lr})$. Due to these relationships, we can now identify conditions under which the value function is increasing on $(\bar{X}(I) \leq_{lr})$ for any $\nu \in \mathbb{R}_0^+$.

Lemma 5.5. *Let (A1) be valid and, additionally, assume $k(\cdot) \equiv 0$ and $c(a_T) \geq c(a_I)$. Then, it holds that $V(\cdot,w,\nu)$ is increasing on $(\bar{X}(I), \leq_{lr})$ for any $w \in \mathbb{N}_0, \nu \in \mathbb{R}_0^+$.*

Proof. Let $x, y \in \bar{X}(I)$ be such that $x \geq_{lr} y$. We show by induction on t that $v_t(\cdot,w,\nu)$ is increasing on $(\bar{X}(I), \leq_{lr})$. Therefore, we assume that $v_{t-1}(\cdot,w,\nu)$ is already increasing on $(\bar{X}(I), \leq_{lr})$. Firstly, we focus on action a_W:
It holds that $\tilde{r}((x,w,\nu),a_W) \geq \tilde{r}((y,w,\nu),a_W)$. By assumption (A1), we have $T(x,a_W,\theta) = x$ and $T(y,a_W,\theta) = y$ for any $\theta \in I\backslash\{N\}$. If $\theta = N$, we obtain $T(x,a_W,\theta) = T(y,a_W,\theta) = e_N^N$. Due to $k(\cdot) \equiv 0$, $v_{t-1}(T(\cdot,a_W,\theta), \zeta((\cdot,w),a_W),\nu)$ then is increasing on $(\bar{X}(I), \leq_{lr})$. Since additionally, it holds that $\rho^{a_W}(x,N) = \rho^{a_W}(y,N)$, we get

$$Lv_{t-1}((x,w,\nu),a_W) \geq Lv_{t-1}((y,w,\nu),a_W). \tag{3}$$

As the optimal operational plan follows an at most 3-action region structure (see Proposition 5.1), we can restrict our analysis to the following cases.

i) $f^{OP*}(x) = f^{OP*}(y) = a_W$:
Due to the penalty costs, for each optimal decision rule $f_t^* \in F$ we have $f_t^*(x,w,\nu) = a_W$, $f_t^*(y,w,\nu) = a_W$. By the help of (3), we obtain

$$v_t(x,w,\nu) \geq v_t(y,w,\nu).$$

ii) $f^{OP*}(x) = f^{OP*}(y) = a_T$:
Due to the penalty costs, for each optimal decision rule $f_t^* \in F$ we have $f_t^*(x, w, \nu)$, $f_t^*(y, w, \nu) \in \{a_W, a_T\}$. Because of $\kappa((x, \nu), a_T) = \kappa((y, \nu), a_T) = 0$, it follows that $\tilde{r}((x, w, \nu), a_T) \geq \tilde{r}((y, w, \nu), a_T)$. As $T(x, a_T, \theta)$ is monotone increasing in x and θ (cf. Lemma 1.2 in Lovejoy, 1987), $T(x, a_T, \theta), T(y, a_T, \theta) \in \bar{X}(I)$ if $\theta \in I \backslash \{N\}$, and $T(x, a_T, N) = T(y, a_T, N) = e_N^N$ otherwise, by using (3) we obtain

$$v_t(x, w, \nu) \geq v_t(y, w, \nu).$$

iii) $f^{OP*}(x) = f^{OP*}(y) = a_I$:
Due to the penalty costs, for each optimal decision rule $f_t^* \in F$ we have $f_t^*(x, w, \nu)$, $f_t^*(y, w, \nu) \in \{a_W, a_I\}$. Because of $\kappa((x, \nu), a_I) = \kappa((y, \nu), a_I) = 0$, it follows that $\tilde{r}((x, w, \nu), a_I) \geq \tilde{r}((y, w, \nu), a_I)$. As $T(x, a_I, \theta) = T(y, a_I, \theta) = e_N^N$, we obtain

$$v_{t-1}(T(x, a_I, \theta), 0, \nu) = v_{t-1}(T(y, a_I, \theta), 0, \nu)$$

and finally, by the help of (3), we get

$$v_t(x, w, \nu) \geq v_t(y, w, \nu).$$

iv) $f^{OP*}(y) = a_W, f^{OP*}(x) \in \{a_T, a_I\}$:
Due to the penalty costs, for each optimal decision rule $f_t^* \in F$ we have $f_t^*(y, w, \nu) = a_W$. By the help of (3), we obtain

$$v_t(x, w, \nu) \geq v_t(y, w, \nu).$$

v) $f^{OP*}(y) = a_T, f^{OP*}(x) = a_I$:
According to Lemma 5.2, it holds that $V^{OP}(x) \geq V^{OP}(y)$. Together with Lemma 5.4, this implies $V(x, w, 0) \geq V(y, w, 0)$ as well as $LV((x, w, 0), a_I) \geq LV((y, w, 0), a_T)$. Due to $c(a_T) \geq c(a_I)$, we get

$$\begin{aligned} \frac{\partial LV((x, w, \nu), a_I)}{\partial \nu} &= -c(a_I) \\ &\geq -c(a_T) \\ &\geq \frac{\partial LV((y, w, \nu), a_T)}{\partial \nu}. \end{aligned} \tag{4}$$

Using inequation (4), we can conclude that $LV(x, w, \nu), a_I) \geq LV((y, w, \nu), a_T)$ for any $\nu \in \mathbb{R}_0^+$. Because of $k(\cdot) \equiv 0$, we have $LV((x, w, \nu), a_I) = Lv_{t-1}((x, w, \nu), a_I)$ as well as $LV((y, w, \nu), a_T) \geq Lv_{t-1}((y, w, \nu), a_T)$ for any $t \in \mathbb{N}$. Finally, we obtain $Lv_{t-1}((x, w, \nu), a_I) \geq Lv_{t-1}((y, w, \nu), a_T)$.

Due to the penalty costs, for each optimal decision rule $f_t^* \in F$ we have $f_t^*(y, w, \nu) \in \{a_W, a_T\}$. Assuming $f_t^*(y, w, \nu) = a_T$, we obtain

$$v_t(x, w, \nu) \geq v_t(y, w, \nu).$$

If $f_t^*(y, w, \nu) = a_W$, using (3) we obtain

$$v_t(x, w, \nu) \geq v_t(y, w, \nu).$$

Finally, this yields that $v_t(\cdot, w, \nu)$ is increasing on $(\bar{X}(I), \leq_{lr})$ for any $t \in \mathbb{N}_0$. □

Using the monotonicity properties of the value function V, we can derive various structural results for the priority index J.

Theorem 5.1. *Assume (A1). Then, it holds that $J(x, w) \geq J(x, w')$ for any $(x, w), (x, w') \in \mathfrak{P}$, where $w \geq w'$.*

Proof. Firstly, we have $f^{OP*}(x) = a \in \{a_T, a_I\}$ for the given optimal operational plan. If $J(x, w) = J(x, w')$, the result is obviously valid. Therefore, we hereafter assume $J(x, w) \neq J(x, w')$. Furthermore, we set $\nu := \min\{J(x, w), J(x, w')\}$ and assume $\nu = J(x, w)$, which is equivalent to $\nu < J(x, w')$. This yields

$$LV((x, w, \nu), a) - LV((x, w, \nu), a_W) = 0. \tag{5}$$

On the other hand, using Lemma 5.3 we obtain

$$LV((x, w, \nu), a_W) - k(w') + k(w) \leq LV((x, w', \nu), a_W)$$

as well as

$$\begin{aligned} &LV((x, w, \nu), a) - LV((x, w, \nu), a_W) \\ &\quad = LV((x, w', \nu), a) + k(w') - k(w) - LV((x, w, \nu), a_W) \\ &\quad \geq LV((x, w', \nu), a) - LV((x, w', \nu), a_W). \end{aligned} \tag{6}$$

Due to the expressions (5) and (6), we immediately get

$$LV((x, w', \nu), a) - LV((x, w', \nu), a_W) \leq 0.$$

Using Lemma 4.1, we finally obtain $\nu \geq J(x, w')$, which contradicts the original assumption $\nu < J(x, w')$. Hence, it holds that $J(x, w) \geq J(x, w')$. □

Theorem 5.2. *Let (A1) be valid and, additionally, assume $k(\cdot) \equiv 0$ and $c(a_T) \geq c(a_I)$. Then, it holds that $J(x, w) \geq J(y, w)$ for any $(x, w), (y, w) \in \mathfrak{P}$, where $x \geq_{lr} y$.*

Proof. Firstly, we have $f^{OP*}(x) = a_1 \in \{a_T, a_I\}$ and $f^{OP*}(y) = a_2 \in \{a_T, a_I\}$ for the given optimal operational plan. If $J(x, w) = J(y, w)$, the result is obviously valid. Therefore, we hereafter assume $J(x, w) \neq J(y, w)$. Furthermore, we set $\nu := \min\{J(x, w), J(y, w)\}$ and assume $\nu = J(x, w)$, which is equivalent to $\nu < J(y, w)$. This implies, that in state (x, w, ν) the decision-maker is indifferent between selecting a_W and a_1. It then holds that

$$LV((x, w, \nu), a_1) - LV((x, w, \nu), a_W) = 0. \tag{7}$$

This implies $V(x, w, \nu) = LV((x, w, \nu), a_1)$. Due to the assumption $\nu < J(y, w)$, we additionally have $V(y, w, \nu) = LV((y, w, \nu), a_2)$. Using Lemma 5.5, we obtain $V(x, w, \nu) \geq$

$V(y, w, \nu)$. If in state (x, w, ν) or (y, w, ν) action a_W is selected and (A1) is valid, the belief state remains the same, unless it transitions into the absorbing state e_N^N. As we assume $f^{OP*}(x) = a_1$ and $f^{OP*}(y) = a_2$, in state (x, w, ν) it is optimal to either select a_W in each future period or to select a_1 after $\tau_1 \in \mathbb{N}_0$ additional periods, unless the project has transitioned into the absorbing state by that time. The same arguments hold for state (y, w, ν). Altogether, we can conclude that, after executing a_W in state (x, w, ν) (or (y, w, ν)) action a_1 (a_2) will be selected after $\tau_1 \in \mathbb{N}_0 \cup \infty$ $(\tau_2 \in \mathbb{N}_0 \cup \infty)$ additional periods. This argumentation leads to the following inequation:

$$\begin{aligned}
&LV((y, w, \nu), a_2) - LV((y, w, \nu), a_W) \\
&\quad = LV((y, w, \nu), a_2) - (\beta(1 - \varpi))^{\tau_2} LV((y, w + \tau_2, \nu), a_2) \\
&\quad \leq LV((y, w, \nu), a_2) - (\beta(1 - \varpi))^{\tau_1} LV((y, w + \tau_1, \nu), a_2) \\
&\quad = \ LV((y, w, \nu), a_2) - (\beta(1 - \varpi))^{\tau_1} LV((y, w, \nu), a_2) \\
&\quad = V(y, w, \nu) - (\beta(1 - \varpi))^{\tau_1} V(y, w, \nu) \\
&\quad \leq V(x, w, \nu) - (\beta(1 - \varpi))^{\tau_1} V(x, w, \nu) \\
&\quad = LV((x, w, \nu), a_1) - (\beta(1 - \varpi))^{\tau_1} LV((x, w, \nu), a_1) \\
&\quad = LV((x, w, \nu), a_1) - LV((x, w, \nu), a_W).
\end{aligned} \tag{8}$$

Due to the expression (7) and (8), we immediately get

$$LV((y, w, \nu), a_2) - LV((y, w, \nu), a_W) \leq 0.$$

Using Lemma 4.1, we finally obtain $\nu \geq J(y, w)$, which contradicts the original assumption $\nu < J(y, w)$. Hence, it holds that $J(x, w) \geq J(y, w)$. □

We can now capitalize on the results of Theorem 5.1 and Theorem 5.2 to establish analog monotonicity results for the critical index. Readopting our superscript notation, we can state the following Corollary.

Corollary 5.1. *Let $b \in B$. The critical index Λ^b in state $(\mathbf{x}, \mathbf{w}) \in \left(\times_{b \in B} \mathfrak{P}^b\right)$ has the following properties:*

i) If (A1) is valid for each $b' \neq b$ and $\mathbf{w}' \in \mathbb{N}^M$ is such that $w'^{b'} \geq w^{b'}$ for each $b' \neq b$, then it holds that

$$\Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b}) \leq \Lambda^b(\mathbf{x}^{-b}, \mathbf{w}'^{-b}).$$

ii) Let (A1) be valid and assume $k^{b'}(\cdot) \equiv 0$ and $c^{b'}(a_T) \geq c^{b'}(a_I)$ for each $b' \neq b$. If $y \in \left(\times_{b \in B} \bar{X}(I^b)\right)$ is such that $y^{b'} \geq_{lr} x^{b'}$ for each $b' \neq b$, then it holds that

$$\Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b}) \leq \Lambda^b(\mathbf{y}^{-b}, \mathbf{w}^{-b}).$$

Proof. The results immediately follow from Theorem 5.1 and Theorem 5.2. □

Finally, we are in a position to provide a well-structured index-based allocation plan covering each state $(\mathbf{x}, \mathbf{w}) \in \mathcal{X}$. On the one hand we know that, assuming (A1) for each $b \in B$, it is optimal to select a_W if $f^{OP,b*}(x^b) = a_W$. On the other hand, we can provide an allocation decision for any state $(\mathbf{x}, \mathbf{w}) \in (\times_{b\in B}\mathfrak{P}^b)$ using the proposed index-based allocation rule. If these two perspectives are combined, we obtain an index rule $f^I = (f^{I,1}, \ldots, f^{I,M})$ which selects the optimal action $f^{I,b}(\mathbf{x}, \mathbf{w}) = a_W$, if $f^{OP,b*}(x^b) = a_W$, and action $f^{I,b}(\mathbf{x}, \mathbf{w}) = f^b(\mathbf{x}, \mathbf{w})$, if $(x^b, w^b) \in \mathfrak{P}^b$.

When only considering states $(x^b, w^b) \in \mathfrak{P}^b$, we can conclude from Corollary 5.1 i) and Theorem 5.1 that the priority index of a project and, thus, the selected action monotonically depend on the waiting time. This means that the considered project is more likely to be operated by a_W as the waiting times of competing projects increase. Correspondingly, it becomes the more likely that the project is operated through $f^{TP,b*}(x^b)$, the more its own waiting time increases. Assuming that the project finally features the largest priority index when its waiting time becomes very large, we can infer the concept of a critical waiting time below which a_W and above which the optimal operational action should be executed. Analogously, Corollary 5.1 ii) and Theorem 5.2 imply that under the given assumptions the proposed allocation rule f^I partitions $\bar{X}(I^b)$ into a set of belief states where a_W is optimal and a set of belief states where $f^{OP,b*}(x^b)$ is optimal. Referring to this, action $f^{OP,b*}(x^b)$ is selected by f^I in any belief state $x^b \in \bar{X}(I^b)$ where $f^{OP,b*}(x^b) \in \{a_T, a_I\}$ and $J^b(x^b, w^b) \geq \Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b})$. For these belief states we can readily transfer the structure of $f^{OP,b*}$ regarding the action regions of a_T and a_I to the allocation rule f^I. Similar to Proposition 5.1, we then obtain two threshold belief states $\mu_1(\cdot)$ and $\mu_2(\cdot)$ which divide each interval into at most three segments.

Proposition 5.2. *Let $b \in B$ and assume (A1) to be true for each project. Additionally, assume $k^b(\cdot) \equiv 0$ and $c^b(a_T) \geq c^b(a_I)$. Then, the index-based allocation rule f^I divides each segment $[y^b, z^b]$ in at most three segments, where $\mathbf{y}, \mathbf{z} \in \left(\times_{b=1}^M \bar{X}(I^b)\right)$ are such that $z^b \geq_{lr} y^b$. Consider $(\mathbf{x}, \mathbf{w}) \in \mathcal{X}$, where $x^b \in [y^b, z^b]$ and $z^{b'} = x^{b'}$ for each $b' \neq b$. Then, there exist $\mu_1^b(\mathbf{x}^{-b}, \mathbf{w}), \mu_2^b(\mathbf{x}^{-b}, \mathbf{w}) \in [y^b, z^b]$ ($y^b \leq_{lr} \mu_1^b(\mathbf{x}^{-b}, \mathbf{w}) \leq_{lr} \mu_2^b(\mathbf{x}^{-b}, \mathbf{w}) \leq_{lr} z^b$) such that*

$$f^{I,b}(\mathbf{x}, \mathbf{w}) = \begin{cases} a_W & , y^b \leq_{lr} x^b \leq_{lr} \mu_1^b(\mathbf{x}^{-b}, \mathbf{w}), \\ a_T & , \mu_1^b(\mathbf{x}^{-b}, \mathbf{w}) \leq_{lr} x^b \leq_{lr} \mu_2^b(\mathbf{x}^{-b}, \mathbf{w}), \\ a_I & , \mu_2^b(\mathbf{x}^{-b}, \mathbf{w}) \leq_{lr} x^b \leq_{lr} z^b. \end{cases}$$

Furthermore, the following properties hold:

a) Let $\mathbf{w}' \in \mathbb{N}_0^M$ be such that $w'^{b'} \leq w^{b'}$ for each $b' \neq b$. Then, it follows that

$$\mu_1^b(\mathbf{x}^{-b}, \mathbf{w}') \leq_{lr} \mu_1^b(\mathbf{x}^{-b}, \mathbf{w}),$$
$$\mu_2^b(\mathbf{x}^{-b}, \mathbf{w}') \leq_{lr} \mu_2^b(\mathbf{x}^{-b}, \mathbf{w}).$$

b) Let $\mathbf{s} \in \left((\times_{b\in B}\bar{X}(I^b)\right)$ be such that $x^{b'} \geq_{lr} s^{b'}$ for $b' \neq b$, if $k^{b'}(\cdot) \equiv 0$ and $c^{b'}(a_T) \geq c^{b'}(a_I)$ are valid for b', and $x^{b'} = s^{b'}$ otherwise. Then, it follows that

$$\mu_1^b(\mathbf{s}^{-b}, \mathbf{w}) \leq_{lr} \mu_1^b(\mathbf{x}^{-b}, \mathbf{w}),$$
$$\mu_2^b(\mathbf{s}^{-b}, \mathbf{w}) \leq_{lr} \mu_2^b(\mathbf{x}^{-b}, \mathbf{w}).$$

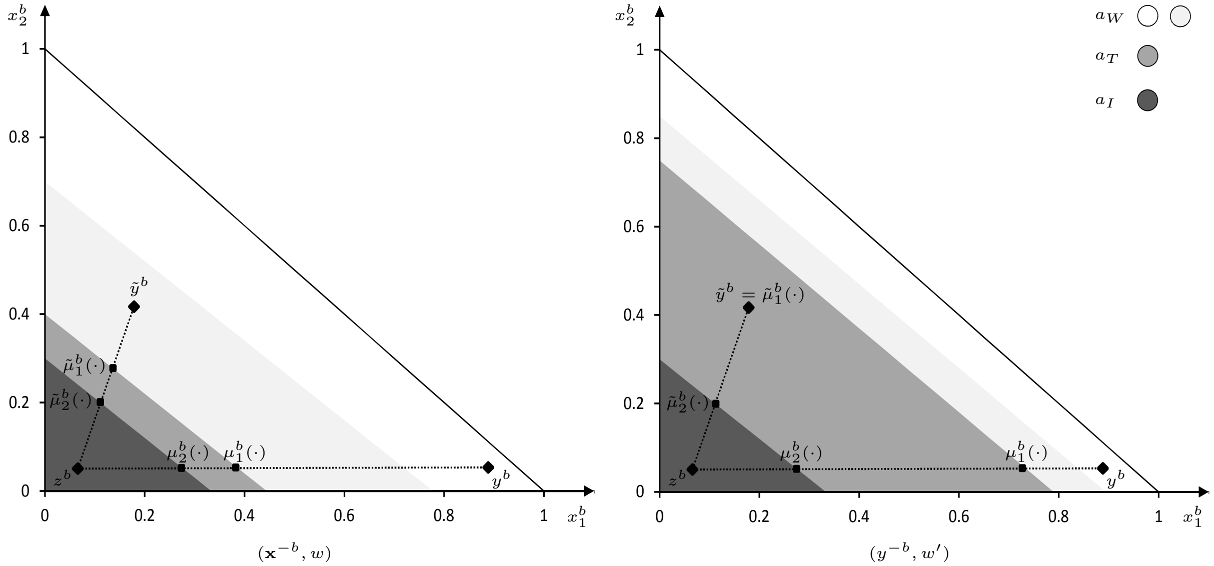

Figure 2: Schematic illustration of the decision rule presented in Proposition 5.2.

Proof. The result immediately follows from Theorem 5.1,Theorem 5.2, Corollary 5.1 and the above given argumentation. □

In Figure 2, we schematically illustrate the proposed allocation rule for an individual project, where $N^b = 4$. The allocation rule $f^{I,b}$ only deviates from $f^{OP,b*}$ if the optimal active action cannot be executed since other projects have a larger priority index. If project b has got the largest index value and, thus, the highest priority, the action selected by $f^{I,b}$ corresponds to the action proposed by $f^{OP,b*}$. If the priority of the considered project decreases, the set of states where a_W is selected, although action $f^{OP,b*}(x^b) \neq a_W$ is optimal, - displayed by the light gray area - monotonically increases with respect to $\leq_{lr}$ such that an at most 3-action region structure results. Consider the two scenarios $(\mathbf{x}^{-b}, \mathbf{w})$, $(\mathbf{y}^{-b}, \mathbf{w}') \in (\times_{b' \neq b} X(I^{b'})) \times \mathbb{N}_0^M$, where $x^{b'} \geq_{lr} y^{b'}$ and $w^{b'} > w'^{b'}$ for each $b' \neq b$. By Proposition 5.2 the critical index $\Lambda^b(\mathbf{x}^{-b}, \mathbf{w}^{-b})$ is larger than $\Lambda^b(\mathbf{y}^{-b}, \mathbf{w}'^{-b})$. The light gray region, therefore, is larger in scenario $(\mathbf{x}^{-b}, \mathbf{w})$ than in $(\mathbf{y}^{-b}, \mathbf{w}')$.

The presented allocation rule again meets our goal to provide a simple, intuitive, and realistic prioritization scheme to a large extent. The index rule allocates the available resources based on project-specific priority indices and, after having determined the projects' order, resorts to an at most 3-action region rule. Moreover, the priority index monotonically depends on the state components of the considered project. Finally, the influence of a project on the operational decisions for other projects can be expressed by monotone relationships.

6 Conclusion

The problem of the optimal allocation of scarce resources to evolving projects is of fundamental importance in many application areas such as the maintenance of a machinery, the allocation of resources to competing research projects or the management of a health care facility. In particular, the trade-off between partially observable project development, waiting times and resource requirements is a wide-spread problem.

In this paper, we model a partially observable allocation scenario drawing on a dynamic optimization approach. Based on the uncertain project stages and the current waiting times, it must decided how the projects should be operated in order to maximize the expected total reward without exceeding the resource capacity. As the allocation problem turns out to be computationally intractable, we propose a heuristic allocation rule based on the well-known MAB approach for solving it.

By incorporating both the problem which project should be activated and the problem which operation mode should be selected, our model allows a comprehensive analysis of how the various factors influence each other. To increase the understanding of the model's recommendations and to simplify the implementation, the focus of our study lies on the identification of structural properties. Specifically, we are able to prove that under the assumption of stable projects the optimal operational plan is characterized by an at most 3-action region rule. Furthermore, we are able to show that the proposed index rule forms a monotone switching curve. Combining these results, we obtain a highly structured allocation rule, i.e. there again results an at most 3-action region rule, which is monotone in the state of each other project.

References

Ansell, P. S., K. D. Glazebrook, J. Nio-Mora and M. O'Keeffe (2003). Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, **57(1)**, 21–39.

Archibald, T. W., D. P. Black and K. D. Glazebrook (2009). Indexability and Index Heuristics for a Simple Class of Inventory Routing Problems. *Operations Research*, **57(2)**, 314–326.

Blackwell, D. (1965). Discounted Dynamic Programming. *The Annals of Mathematical Statistics*, **36(1)**, 226–235.

Gittins, J. and D. Jones (1979). A dynamic allocation index for the discounted multi-armed bandit problem. *Biometrika*, **66(3)**, 561–565.

Gittins, J. C., K. D. Glazebrook and R. R. Weber (2011). *Multi-armed Bandit Allocation Indices.* John Wiley & Sons, London.

Glazebrook, K. D. (1982). On a Sufficient Condition for Superprocesses Due to Whittle. *Journal of Applied Probability*, **19(1)**, 99–110.

Glazebrook, K. D. and R. Minty (2009). A Generalized Gittins Index for a Class of Multiarmed Bandits with General Resource Requirements. *Mathematics of Operations Research*, **34(1)**, 26–44.

Glazebrook, K. D., H. M. Mitchell and P. S. Ansell (2005). Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, **165(1)**, 267–284.

Jin, L., T. Mashita and K. Suzuki (2005). An optimal policy for partially observable Markov decision processes with non-independent monitors. *Journal of Quality in Maintenance Engineering*, **11(3)**, 228–238.

Krishnamurthy, V. and B. Wahlberg (2009). Partially Observed Markov Decision Process Multi-armed Bandits - Structural Results. *Mathematics of Operations Research*, **34(2)**, 287–302.

Le Ny, J., M. Dahleh and E. Feron (2008). Multi-UAV Dynamic Routing with Partial Observations using Restless Bandit Allocation Indices. In *Proceedings of the American Control Conference*, 4220–4225.

Lewis, M. E. (2001). Average Optimal Policies in a Controlled Queueing System with Dual Admission Control. *Journal of Applied Probability*, **38(2)**, 369–385.

Lovejoy, W. S. (1987). Some Monotonicity Results for Partially Observed Markov Decision Processes. *Operations Research*, **35(5)**, 736–743.

Mahajan, A. and D. Teneketzis (2008). Multi-Armed Bandit Problems. In Hero, A. O., D. A. Castan, D. Cochran and K. Kastella (eds.), *Foundations and Applications of Sensor Management*, 121–151. Springer, New York.

Monahan, G. E. (1980). Optimal Stopping in a Partially Observable Markov Process with Costly Information. *Operations Research*, **28(6)**, 1319–1334.

Ohnishi, M., H. Kawai and H. Mine (1986). An optimal inspection and replacement policy under incomplete state information. *European Journal of Operational Research*, **27(1)**, 117–128.

Papadimitriou, C. H. and J. N. Tsitsiklis (1999). The Complexity of Optimal Queuing Network Control. *Mathematics of Operations Research*, **24(2)**, 293–305.

Sondik, E. J. (1978). The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs. *Operations Research*, **26(2)**, 282–304.

Weber, R. R. and G. Weiss (1990). On an Index Policy for Restless Bandits. *Journal of Applied Probability*, **27(3)**, 637–648.

WHITT, W. (1979). A Note on the Influence of the Sample on the Posterior Distribution. *Journal of the American Statistical Association*, **74(366)**, 424–426.

WHITTLE, P. (1981). Arm-Acquiring Bandits. *The Annals of Probability*, **9(2)**, 284–292.

WHITTLE, P. (1988). Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, **25**, 287–298.

ZHELTOVA, L. (2010). *Structured Maintenance Policies on Interior Sample Paths.* Phd thesis, Case Western Reserve University.