Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Learning in finance

Master's thesis

2013 Bc. Vladimír Novák

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics
Department of Applied Mathematics and Statistics

Learning in finance

Master's thesis

Bc. Vladimír Novák

Supervisor:
prof. RNDr. Pavel Brunovský, DrSc.

Branch of study: 1114 Applied Mathematics
Study programme: Economic and Financial Mathematics

BRATISLAVA 2013

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky
Katedra aplikovanej matematiky a štatistiky

Modely učenia vo financiách

Diplomová práca

Bc. Vladimír Novák

Školiteľ:
prof. RNDr. Pavel Brunovský, DrSc.

Študijný odbor: 1114 Aplikovaná matematika
Študijný program: Ekonomická a finančná matematika

BRATISLAVA 2013

60793109

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Vladimír Novák
Study programme:	Economic and Financial Mathematics (Single degree study, master II. deg., full time form)
Field of Study:	9.1.9. Applied Mathematics
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Learning in finance

Aim: The goal of the thesis is to explore the possibilities of the application of multi-armed bandit theory to models of venture capitalists investments with learning.

Supervisor:	prof. RNDr. Pavel Brunovský, DrSc.
Department:	FMFI.KAMŠ - Department of Applied Mathematics and Statistics
Vedúci katedry:	prof. RNDr. Daniel Ševčovič, CSc.
Assigned:	25.01.2012

Approved: 26.01.2012 prof. RNDr. Daniel Ševčovič, CSc.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Vladimír Novák
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Modely učenia vo financiách

Cieľ: Cieľom práce je preskúmať možnosti aplikácie "multi-armed bandit" teórie na modely rizikového investovania s učením.

Vedúci: prof. RNDr. Pavel Brunovský, DrSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 25.01.2012

Dátum schválenia: 26.01.2012 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Learning in finance

Bc. Vladimír Novák
E-mail: *novakvlado@gmail.com*
Web-page: `http://sk.linkedin.com/pub/vladimír-novák/66/638/63a/`

prof. RNDr. Pavel Brunovský, DrSc.
E-mail: *brunovsky@fmph.uniba.sk*
Web-page: `http://www.iam.fmph.uniba.sk/institute/brunovsky/`

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
Mlynská dolina, 846 48 Bratislava
Slovakia

Master's thesis in Applied Mathematics
Compilation date: April 23, 2013
Typeset in LaTeX

Abstrakt

Bc. Vladimír Novák: Modely učenia vo financiách [Diplomová práca].
Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky,
Katedra aplikovanej matematiky a štatistiky.
Školitel: prof. RNDr. Pavel Brunovský, DrSc.
Bratislava 2013

V práci sa zaoberáme možnosťami opísať investovanie investorov s rizikovým kapitálom pomocou metódy "multi-armed restless bandits". Použitím klasickej verzie "multi-armed bandits" na odvodenie učiaceho sa modelu Sorensen (2008) ukázal, že učenie sa investorov o neistotách spojených s investičnými možnosťami a technológiami je pri ich investíciách bežné. Okrem toho Sorensen jasne zamietol hypotézu, že individuálne investície sa robia v izolácii. Formulácia problému pomocou "restless" verzie "multi-armed bandits" nám umožňuje vyhnúť sa nevyhovujúcim predpokladom zo Sorensenovho článku. V práci predstavujeme tri modely opisujúce investovanie rizikového kapitálu v spomínanom prostredí a pre všetky z nich odvodzujeme Whittlovu indexovú stratégiu.

Na základe jedného z uvedených modelov vytvárame učiaci sa model používajúci Bayesovské aktualizovanie a formulujeme ho v prostredí čiastočne pozorovateľných Markovovských rozhodovacích procesov. Numericky počítame Whittlovu indexovú stratégiu pre tento učiaci sa model. Taktiež uvádzame výsledky reprezentatívnej vzorky numerických simulácií na ohodnotenie výkonnosti indexu voči zvyčajnej stratégii odhadu návratnosti investície. Táto simulačná štúdia ukazuje, že naše riešenie funguje dobre a spravidla prekonáva výkonnosť všeobecne používaného riešenia pomocou návratnosti investície.

Klúčové slová: Multi-armed restless bandit • Bayesovské učenie • Markovovské rozhodovacie procesy • Indexové stratégie • Rizikový kapitál

Abstract

Bc. Vladimír Novák: Learning in Finance [Master's thesis].
Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics,
Department of Applied Mathematics and Statistics.
Supervisor: prof. RNDr. Pavel Brunovský, DrSc.
Bratislava, 2013

In this thesis we investigate possibilities to capture the problem of venture capitalists (VCs) investments in entrepreneurial companies by the multi-armed restless bandits framework. As shown in Sorensen (2008), by adoption of the classical multi-armed bandits model for deriving the learning model, VCs' learning about investment opportunities and technology uncertainties is prevalent for their investments. Moreover, the hypothesis that individual investments are done in isolation is clearly rejected by Sorensen. Formulation of the problem by the restless version of the multi-armed bandits allows us to avoid not fully reasonable assumptions for the financial applications from the Sorensen's paper. We provide three different models in this methodology for describing the VCs investments and we derive the Whittle's index policy for all of them.

Based on one of these models we develop a learning model which incorporates the Bayesian updating and we formulate the model as a partially observable Markov decision process. We numerically obtain the Whittle's index policy for the learning model. We also report on a number of numerical simulations for the index performance evaluation against the usually used return on investment approach. This simulation study suggests that our solution is well performing and often outperforms the return on investment generally employed solutions.

Keywords: Multi-armed restless bandit • Bayesian updating • Markov decision process • Index policies • Venture capital

Acknowledgements

I would like to express special thanks to my supervisor prof. RNDr. Pavel Brunovský, DrSc. for all the support, guidance and corrections of my writings.

My special thanks goes to Mgr. Peter Jacko, PhD. and Sofia Villar, PhD. for the friendship, help and invaluable advices for research and life they offered me throughout my bachelor and master studies. I would also like to thank prof. Ľuboš Pástor for initial inspiration and help.

Last but not least, I warmly thank my brother Marek and Veronika for support and love.

This master thesis was supported by a grant from Nadácia Tatrabanky, scholarship program Hlavička of Nadácia SPP and my master studies were supported by U. S. Steel Košice.

Declaration on Word of Honour

I declare on my honour that this thesis was written on my own, with the only help provided by my supervisor and the referred-to literature and sources.

In Bratislava April 23, 2013 .....

Bc. Vladimír Novák

Contents

List of Figures

List of Tables

Introduction

The most confusing moments in our lives are usually connected to decision making. In these situations it is scarce to have a full information, therefore we have to deal with the phenomenon of *uncertainty*. This is one of the reasons why it is so difficult to make a rational decision. When we realize that we are able to learn about the uncertain parameters, we can solve many life problems that appear puzzling at first sight more easily.

Once we developed our beliefs about the parameters, we can deal with a decision problem by setting priorities to each alternative and choosing the alternative with the highest priority. Such tasks arise in all fundamental economic problems where we have to allocate scarce resources to a number of alternative uses. Therefore, it is of a great practical interest to develop a methodology for establishing suitable priorities to different alternatives. In the presence of uncertainty we not only have to sacrifice the benefits of the unselected alternatives, but also the information provided by the unselected ones. For instance, we consider several entrepreneurial companies competing for the available investment at the same time. Suppose that independently of other companies, we can associate a value with each company. This value determines the efficiency of attaining a joint goal if we allocate resources to it at a given moment. We refer to this value as an *index*. In addition, we also need to take into consideration the consequences of an ubiquitous phenomena: *bankruptcy*.

From the mathematical point of view, such problems could be formulated as discrete-time *Markov decision processes* and solved by employing recent developments of the theory of *Multi-armed restless bandits* for deriving a simple implementable scheduling rule (proposed by Whittle (1988)). This scheduling rule is based on assigning an index to every company and investing in the company with the highest priority. In this thesis our objective is to solve an example of such problem in the presence of uncertainty. More specifically, we focus on financial problems, because the financial markets are naturally connected with a large amount of randomness and thus agents have to learn about parameters characterizing financial markets by observing data. An overview of such problems can be found in the paper from Pastor and Veronesi (2009) that reviews recent work on learning in finance, especially

applications related to the portfolio choice, stock price bubbles, mutual fund flows, trading volume, etc.

Investing by *Venture capitalists (VCs)* in entrepreneurial companies is a suitable example of the above mentioned problem. Investors are uncertain about technologies and investment opportunities. While there was a surge of papers dealing with relationship between VCs and their entire portfolios (see for example Hochberg et al. (2007)), less is known about their particular investments (exception for example: Kaplan and Stromber (2004)), which were also shown to be filled by uncertainty (Quindlen (2000)). VCs learning is essential for understanding their investment decisions. Sorensen (2008) by his econometric study showed that VCs investment decision is based on the expected return from the investment itself and on the potential to learn from it. He also rejected the hypothesis that VCs' investments are chosen independently to maximize the return from each investment individually, as it is predicted by standard models.

In order to develop the VCs learning model, Sorensen extended the *classical* multi-armed bandit problem (see Gittins (1989)). The latter is a stochastic and dynamic resource allocation model with special structure, specifically it is a model of a controller optimizing her decisions while acquiring knowledge at the same time. Originally it was inspired by a gambler problem, how to select which slot machine (a.k.a. one-armed bandit) she should play in casino. Bandits incorporation has two advantages. Firstly, it allows us to distinguish between the influence of investor's learning from the past investments (*exploitation*) and the option value of future learning (*exploration*), when making investment decisions. This optimal strategy trades off between the investments for profit and the investments for learning. This is one of the biggest contributions of Sorensen's paper. Exploitation investments have high known payoffs and exploration investments have uncertain payoffs, but they often provide higher option value of learning. Sorensen finds that VCs who learn more are more successful in the long-term.

Secondly, by incorporating the classical multi-armed bandits Sorensen is able to avoid computationally intensive estimation procedures to capture the intractable dynamic programming problem (see for example Crawford and Shum (2005)). Index result of the model helps to simplify the empirical analysis by allowing the model to be estimated using standard statistical procedures.

Sorensen (2008) is the cornerstone for this thesis. Incorporation of the classical multi-armed bandits causes also disadvantages, mostly by requiring additional assumptions that are not entirely reasonable in the context of entrepreneurial investing. The model assumes that investors choose between investments at the industry level. Another assumption is that the environment is stationary. Therefore, investors only learn from their own past investments and investments in one industry are not informative about investments in other industries. Moreover Sorensen's paper does not describe the dynamics of VCs investments, which is another disadvantage we try

to take care of.

The main aim of the thesis is to develop a dynamic learning model describing VCs investment decisions. To avoid previous unreasonable assumptions we use recent developments of the *restless* multi-armed bandits (see Whittle (1988)). On the other hand, it forces us to describe investments dynamics. To the best of our knowledge this is the first time the VCs investments decision is captured by the multi-armed restless bandits framework. To summarize, the main goals of the thesis are:

- To design a dynamic model describing VCs investing in entrepreneurial companies.
- To propose a model based on the multi-armed restless bandits theory incorporating Bayesian updating.
- To avoid the following assumptions in the proposed models:
  - Investors choose between entrepreneurial companies only at the industry level.
  - Investors learn only from their own past investments.
  - Investments in one industry are not informative about investments in other industries.

A contribution of this thesis is also in a unique combination of the mathematical approaches from different areas. This combination is illustrated by the following scheme.

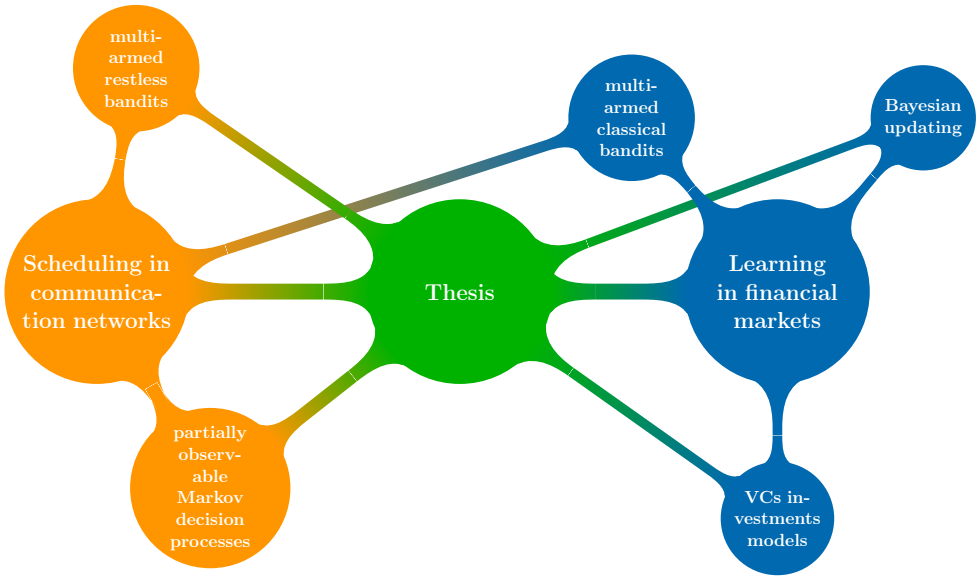

Figure 0.1: Schematic illustration of the mathematical approaches combination

The thesis is organized as follows. The first chapter provides a theoretical background for the part of the thesis not dealing with learning. Those are the classical

multi-armed bandits model, the restless multi-armed bandits model, description of a Markov decision process framework and bandits terminology in the finance environment. In chapter 2 three different models are proposed describing VCs investment decisions with their MDP formulation. Moreover, the index values derivation for all three models is proposed there as well. Chapter 3 provides theoretical background for the part of the thesis dealing with learning. Thus, it consists of Bayesian updating and description of partially observable Markov decision process (POMDP) framework. Finally chapter 4 presents the Bayesian venture capitalists model. We simulate the index solution and show the comparison with other policies.

Chapter 1

Multi-armed bandits and index policies design methods

This chapter provides an introduction to the theoretical frameworks used throughout the thesis. We mainly focus on solution methods for stochastic dynamic programming problems. First we introduce the Markov decision process framework, followed by the evolution from the classical to the restless multi-armed bandits. Next we provide description how to design index policies and a brief history overview. In this chapter we do not deal with learning and uncertainty. These ubiquitous phenomena will be introduced later. This survey is based on Jacko (2009b), Niño-Mora (2010) and Villar (2012). Some parts could also be found in Novak (2011), but in disparity with these surveys, we try to modify the description to refer more to finance.

1.1 Markov decision process framework

By a *Markov decision process (MDP)* we understand a sequential decision process in which information needed to predict the future evolution of a system is contained in the current state of the system and depends on current action. For example, in stochastic and dynamic resource allocation problems, future evolution of the underlying system depends on scheduler's chosen action at various time instants. After the scheduler has selected the action she earns reward and the system evolves in the prescribed way that is action depending. The scheduler wants to minimize the expected total[1] cost or to maximize the expected total reward over a certain time horizon. The horizon can be finite or infinite. In the infinite case we can use *discounting* or *long-run averaging* in order to have a finite-valued objective (Stidham, 2002). In finance applications the scheduler could be an investor trying to maximize an expected total reward over a certain time horizon. His action is to invest or not to invest money in a particular company and based on this decision she earns or loses the money. Due to the financial behaviour of our models, it is mainly the discounting approach for

[1]Throughout the thesis the term total is reserved to mean the sum over all time instants.

the expected total reward which is used in the thesis.

The major strength of the MDP framework lies within its wide modelling power. It is used in a variety of applications such as economics, management science, operations research, applied probability and engineering systems. In this thesis we focus on discrete-time MDP theory. For VCs investing it is natural to be carried out in discrete points of time and not continuously.

In decision making moments we do not usually have any information about future states. Therefore, in MDP decision rules are assumed to be *non-anticipative*, i.e. history-dependent, which is defined as a set of rules specifying the action to be taken for each decision point in time and for each possible state of the system, using only current and past information (Jacko (2009b)). The term *decision rule* specifies the action to be chosen at a particular time and a *policy* is a sequence of decision rules. In other words, it tells us what to do at any time if the system is in a given state. A policy is *stationary* if it is time-homogeneous, so it is not depending on the time instant. MDPs are of Markovian nature (future evolution depends only on current state), thus such policy is appropriate.

Markov decision processes could be studied by *dynamic programming* developed by Richard Bellman in the 1950s. The cornerstone for this method is the *Principle of Optimality: "At any point in time, an optimal policy must prescribe an action that optimizes the sum of immediate reward and expected total reward obtained if an optimal policy is applied from the subsequent point in time on"* (see: Jacko (2009b) and Bellman (1957)). Optimality equations of dynamic programming, known as *Bellman equations*, constitute the mathematical framework developed in pursuance of *Principle of Optimality*. From the Bellman equation we can derive theoretical results as necessary and sufficient condition for optimality of a stationary policy in a variety of cases. Practically it leads to a recursive solution method for dynamic programming problems, significantly decreasing the problem complexity. Nevertheless, this reduction is not sufficient and for many problems the solution is still intractable.

In a number of cases we are confronted with *the curse of dimensionality*. It means that dynamic programming formulations grow exponentially with the number of states variables. This forces us to developed other approaches. One of such approaches is *Lagrangian relaxation*, that helps us to decompose complex problems with special structure to a family of subproblems from which we are able to obtain well-performing suboptimal solutions. The other one is *linear programming* (LP) reformulation, where an optimization term from each Bellman equation can be relaxed to a set of linear inequalities, where each represents exactly one action. The LP approach is suitable to constrained MDPs, in which the optimal policy must satisfy side constraints (Stidham (2002)). Thus in some cases it is possible to solve such problem by LP after reformulation, which should also be possible in our models proposed in the chapter 2. Besides Lagrangian relaxation we use some developments of the LP reformulation in this thesis.

1.2 Classical multi-armed bandits

Multi-armed bandits are named after *one armed bandit slot machine* that one can find in casinos (see Figure 1.1). Obviously the difference between one-armed bandits and multi-armed bandits are in the number of levers the gambler can pull, e.g. if the gambler faces several slot machines, or one slot machine with multitude of levers. The problem is which arm (exactly one at a time) should the gambler pull and in which order. After the gambler pulls a lever she receives random reward from the distribution specific to that particular lever. The objective is to maximize total earned reward after a sequence of trials.

Figure 1.1: One armed bandit - slot machine that one can find in casinos
Source: www.cashcashpinoy.com

This problem was originally introduced by Robbins (1952a), where he constructed convergent population selection strategies. In this model the controller optimizes decisions and receiving new informations at the same time. It can be reformulated as the problem solving dynamic allocation of a single scarce resource amongst several stochastic alternative projects (Weber (1992)). It was quite a challenging stochastic dynamic optimization problem, until Gittins and Jones (1974) proposed an optimal policy for maximizing the expected discounted reward.

In classical multi-armed bandits the played bandit is represented by a random reward yielding process. If not played it stays *frozen*, so no state evolution and no

rewards occur. This is the main difference between classical and restless bandits (introduced later). The problem models balance between getting the highest immediate reward and learning about the system (information about distribution specific to the particular lever) and receiving possibly even higher rewards later. Often it is referred to as trade-off between exploitation and exploration, known in reinforcement learning.

In practice it is used to model the problem of managing research projects in a large organization, like a science foundation or a pharmaceutical company. For instance, investigating the impact of different experimental treatments and minimizing patient losses at the same time.

1.3 Restless multi-armed bandits

The *multi-armed restless bandits problem* proposed by Whittle (1988) represents a generalization of the multi-armed bandits. It added two features to the classical version. Bandits are no longer frozen when they are not played, so they are allowed to evolve and yield reward. The second feature is that we can allocate scarce resources parallely to a fixed number of bandits (Jacko (2009b)). More precisely: *"Multi-armed restless bandits are Markov decision process models for optimal dynamic priority allocation to a collection of stochastic binary-action (active/passive) projects evolving over time"* (Niño-Mora (2010)). Extensions lead to problems with tractability. It was proven that the multi-armed restless bandits are P-SPACE hard, even in the deterministic case (Papadimitriou and Tsitsiklis (1999)) .

1.3.1 Example: Portfolio project

Following Niño-Mora (2010) we introduce here an application to portfolio project problem. Instead of bandits we can imagine dynamic and stochastic projects. Imagine a collection of $N \geq 2$ projects (one-armed bandits) to be labelled by $k = 1, \ldots, N$. At each time instant the manager can choose $M \leq N$ projects on which he wants to work. The projects can be in several states, the state of which we denote by $\mathbb{X}$ and it is same for all the projects.

At the start of each period the manager has an option to work (active) or not to work (passive) on a particular project. His decision on the project k will be represented by $a_k = 1$ if he is active on the project and $a_k = 0$ if not. Dependently on his choice, if the project k is in a state $X_k(t) = i \in \mathbb{X}$ at the start of the time period and $a_k(t) = \delta$, in the restless bandits version it moves with transition probability $p_k(i, j|\delta)$ to state $X_k(t+1) = j \in \mathbb{X}$ and it yields an immediate random reward $R_k(i, \delta)$, where $\delta = \{0, 1\}$. Unlike in the classic bandits version, if $a_k(t) = 0$ the state does not change, i.e. $p_n(i, i|0) = 1$. We incorporate a scalar parameter λ into the model, which represents the charge incurred per active period. Thus the net reward for active action is $R_n(i, 1) - \lambda$.

At the start of each period t the project manager observes the joint state $\mathbf{X}(t) = (X_k(t))_{k=1}^N$ and based on the history of joint states and actions satisfying $\sum_{k=1}^N a_k(t) \leq M$, he takes a joint action $\mathbf{a}(t) = (a_k(t))_{k=1}^N$. The infinite-horizon λ-charge multi-armed restless bandit problem is to find an admissible scheduling policy π^*, which maximizes the expected total discounted net reward. The scheduling policy π^* is a sequence of non-anticipative decision rules (joint actions) $\mathbf{a}(t)$ that prescribes on which project we should work at time t and π^* is chosen from the resulting class $\mathbf{\Pi}(M)$ of all admissible scheduling policies.

We formulate the above mentioned problem as:

$$\max_{\pi \in \mathbf{\Pi}(M)} \mathbf{E}_{i^0}^{\pi} \left[\sum_{t=0}^{\infty} \sum_{k=1}^{N} \{R_k \left(X_k(t), a_k(t)\right) - \lambda a_k(t)\} \beta^t \right],$$

where $\mathbf{E}_{i^0}^{\pi}$ denotes expectation for a fixed initial portfolio state $\mathbf{X}(0) = \mathbf{i}^0 = (i_k^0)_{k=1}^N$ and under policy π.

1.4 Index policies

In the early 1970s, Gittins proposed the concept of index priority policy (also called index rule) for the classical multi-armed bandit problems and proved that it is optimal (see Whittle (1980)). It assigns a dynamic allocation index to each competing bandit and allocate the scarce resources to a bandit with the highest current index value. The index solution is important because it could be evaluated separately for each bandit. This solution for the classical multi-armed bandits is know as the *Gittins priority index policy* and the proposed index is known as *Gittins index* (Gittins (1979)). A very elegant proof of optimality could be found in Weber (1992).

To solve the restless multi-armed bandits we use the index based solution proposed by Whittle (1988). To do so we replace a family of sample-paths by a unique one. In other words, we relax the constraint that we are playing the fixed number of bandits at each time period to be constraint that we are playing the required number of bandits only on average. Using Lagrangian relaxation allows us to decompose the problem into separate subproblems that could be solved separately. The obtained optimal solution for a unique bandit is then used to develop a heuristic rule for the original problem. The scarce resources are again allocated to the bandit with the highest current index value.

In general such index solution to the restless multi-armed bandit problems usually has only some form of asymptotic optimality as was shown by Weber and Weiss (1990). It is often nearly-optimal and better than ad hoc solutions. On the other hand, Whittle (1988) realized that not for all restless bandits an index exists. We call a bandit *indexable* if such index exists for that particular bandit. Methods for analysing bandits indexability were presented in Niño-Mora (2001, 2002) and Niño-

Mora (2006).

Proposed indices often have an economic interpretation (see Jacko (2009b)). For instance, the Gittins index satisfies the maximal reward rate. The reason is that it is the maximal rate of expected rewards per unit of expected time. The index developed by Whittle is characterized as a fair charge for assigning the scarce resource to the bandit. The *MP index* introduced by Niño-Mora (2002, 2006) is a generalization of all above mentioned indices. From the economic perspective it could be described as the marginal rate of transformation of employing a scarce resource at a given state of a bandit.

1.5 Overview of the multi-armed bandits history

Origins of the classical multi-armed bandits could be found in the seminal works by Thompson (1933) and Robbins (1952b) focused on the area of sequential design of experiments. These developments found their applications mainly in the optimal dynamic allocation of patients to clinical treatments with unknown success probabilities. In such cases we can refer to exploitation and exploration. Exploitation is that for the next patient we can use treatment which is the best one from our historical data and exploration can be observed in the opportunity to try a treatment which does not yield such as good immediate improvement, but we have a belief that it could turn to be the best one. On the other hand, officially the classical multi-armed bandits problem was formulated during Second World War by Allied scientists. According to statements of Peter Whittle it was proved to be intractable and passed to German scientists that they also can waste their time by solving this problem.

As we already mentioned these problems are MDPs and thus could be solved by dynamic programming. Unfortunately, dynamic programming does not provide any insight to the structure of these problems and for the restless case we have problems with the curse of dimensionality. This forced researchers to focus on solutions based on the special structure of these problems. Bradt et al. (1956) was the first who showed optimality of the index solution for the classic finite-horizon undiscounted one-armed bandit problem. Extension for infinite-horizon was carried out by Bellman (1956). His index solution was a function of state only. For a long time researchers were trying to apply similar ideas to the classical multi-armed bandits but without any success until Gittins and Jones (1974) proposed their solution known as the *Gittins index*. It received a wide attention and became very popular. Nevertheless, there still was an unsolved important extension were bandits evolve even if not played.

Whittle (1988) realized it and proposed his own heuristic solution for the restless multi-armed bandits. It was based on Lagrangian relaxation and decomposition approach which resulted in index heuristics. He also found out that it holds only for bandits with special structure which he calls indexable. There was a surge of works focused on developing general sufficient conditions for indexability what was

published in papers: Niño-Mora (2001, 2002) and Niño-Mora (2006). Niño-Mora also proposed an adaptive-greedy algorithm for indexability verification.

Nowadays bandits are used for various applications as wireless systems, telecommunications, etc.; but based on our knowledge the restless multi-armed bandits were never used for VCs investing.

1.6 Bandits terminology for finance applications

So far we used general informations about multi-armed bandit problems and also usual notation and terminology. Here we want to introduce a dictionary from the bandits terminology to the financial terminology.

Table 1.1: Dictionary from the general bandits terminology to the financial terminology

Bandits terminology		Financial terminology
scheduler/controller	→	investor, venture capitalist, angel investor
bandit	→	industry
lever	→	entrepreneurial company, company
to play/pull a particular lever	→	to invest into the company from particular industry
epoch	→	instant (decision moment)
slot	→	period (when waiting for outcome)

Chapter 2

Venture capitalists investments into the entrepreneurial companies

As we already mentioned the dynamics of VCs investments into entrepreneurial companies was not captured by the multi-armed restless bandits framework before. Therefore we design three different models describing VCs investing. For simplicity we do not incorporate learning into these models at this point. Each one of them tries to capture a different feature and we can observe differences better if the model is not too complex already. Moreover, the obtained closed-form index solutions could be very helpful and can provide an insight which factors are important for the solution. Formulation of the VCs investments as restless bandits is very important, because restless behaviour allows us to avoid assumption that investments in one company are not informative about investments in another, so when we invest in one company nothing is happening with other companies.

The first model is the simplest thus it requires several restrictive assumptions. We use it mainly to describe the index designing procedure. The second and the third model are trying to avoid other not fully reasonable assumptions from Sorensen (2008). In Novak (2011) we can find the solution of one such models. In this thesis we solve three models the first and the third of which are solved by emulation of *AG-algorithm* that is different from the solution used in Novak (2011). At the end of the chapter we present a summary of the obtained indices and we discuss the applications of the particular models.

2.1 Venture capitalists investments model 1

2.1.1 Problem description model 1

Investors (VCs) can invest in $K-1$ entrepreneurial companies. The opportunity K describes the possibility not to invest. We refer to it as a *alternative investment*. We assume that time is discrete and goes to infinity. At every time instant the investor

chooses exactly one opportunity in which she *invests actively*. An active investment is characterized as the action when the investor does not only hold the company in his portfolio but she actively collaborates with the chosen company and pushes it to *initial public offering (IPO)* or *acquisition*. The same happens when the investor invests in the company for the first time and has to develop a whole new structure etc. Thus active investment causes higher costs for the investor as *passive investment*. During passive investment she only gives necessary money to the company, but she does not do anything in addition.

An investment is characterized by a cost of passive investment $c_k^0 \geq 0$ and by a cost of active investment $c_k^1 \geq 0$, where naturally $c_k^1 \geq c_k^0$. Other parameters characterizing the investment are: success probability $\mu_k \geq 0$ (IPO), bankruptcy probability $\theta_k \geq 0$ and bankruptcy penalty $d_k \geq 0$ describing other losses connected with bankruptcy of the company such as the investor's reputation loss, waste of prepared processes and strategies for that particular company etc. If the investment is successful, the investor gains a reward $\mathcal{R}_k$. We suppose that $\mathcal{R}_k$ is a one time payment, so it is received by the investor only in a time point when the investment succeeds. In reality, the investor earns dividends after IPO or acquisition. Therefore we can look at $\mathcal{R}_k$ as at the present value of an annuity of dividend payments. For active investment into the alternative investment the investor obtains *alternative reward* κ. We assume that the time until success (if active) and the time until bankruptcy (if passive) follow the geometric distribution.

The investor's goal is to maximize the expected aggregate net reward, i.e. aggregate reward minus aggregate investments costs and bankruptcy costs, over an infinite horizon. The investor decides at equidistant time instants, in which company (if any) she should actively invest. In the queueing theory we say that the investor is *preemptive*.

In this model we assume that investors have no budget-related constraint. From the VCs perspective this assumption is suitable. The reason is that VCs usually can borrow a big loan and entrepreneurial companies as start-ups are small investments in the comparison with the possible loan. Moreover, the optimization itself will not allow to go to too big debt positions, since active investment is necessary to get a success.

2.1.2 MDP formulation model 1

We set our discrete-time model without arrivals into the framework of a dynamic and stochastic resource allocation problem and follow Jacko (2009a) approach to design Whittle index policies. The time is partitioned into discrete decision time instants $t \in \mathcal{T} := \{0, 1, 2, \dots\}$, where t corresponds to the beginning of the time period and we refer to it as a decision time point (instant). Suppose that at $t = 0$ there are $K - 1 \geq 2$ entrepreneurial companies waiting for VC investment. The investor at each time instant chooses (at most) one company in which she invests actively. If no

company is chosen, then the investor is allocated to the alternative investment, i.e. there are K competing possibilities, labeled by $k \in \mathcal{K}$. Thus, the investor invests in exactly one option at a time.

Companies and industries model 1

The investor can allocate either zero or full attention to any industry $k = 1, 2, \ldots, K-1$. We denote by $\mathcal{A} := \{0, 1\}$ the *action space*. Action $a = 0$ means that the investor does not actively invest in the company, and action $a = 1$ means that the investor does actively invest in it. This action space is the same for every company k.

Each company/industry k is defined independently of other companies/industries as the tuple

$$\left(\mathcal{N}_k, (\boldsymbol{W}_k^a)_{a\in\mathcal{A}}, (\boldsymbol{R}_k^a)_{a\in\mathcal{A}}, (\boldsymbol{P}_k^a)_{a\in\mathcal{A}}\right),$$

where

- $\mathcal{N}_k := \{*, 0, 1\}$ is the *state space* of the company k, where state $*$ represents a company without any investment, 0 represents a company that had been in investor's portfolio but it either succeeded or bankrupted, and state 1 means that the company is in investor's portfolio (the investor invested actively in the company, but it neither succeeded nor bankrupted).
- $\boldsymbol{W}_k^a := \left(W_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $W_{k,n}^a$ is the (expected) one-period attention consumption, or *work* required by company k at state n if action a is decided at the beginning of a period. In our model it is always the same as the chosen action a; in particular, for any $n \in \mathcal{N}_k$,

$$W_{k,n}^1 := 1, \qquad W_{k,n}^0 := 0;$$

- $\boldsymbol{R}_k^a := \left(R_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $R_{k,n}^a$ is the expected one-period *reward* earned by the investor for company k at state n if action a is decided at the beginning of the period; in particular,

$$\begin{array}{lll} R_{k,*}^1 := -c_k^1, & R_{k,0}^1 := 0, & R_{k,1}^1 := -c_k^1 \cdot (1-\mu_k) + \mathcal{R}_k\mu_k, \\ R_{k,*}^0 := 0, & R_{k,0}^0 := 0, & R_{k,1}^0 := -c_k^0 \cdot (1-\theta_k) - d_k\theta_k \end{array}$$

  Where $\mathcal{R}_k > c_k^1 \geq c_k^0 \geq 0$, $d > c_k^0$ and $\mathcal{R}_k \geq \frac{c_k^1}{\mu_k}$.

- $\boldsymbol{P}_k^a := \left(p_{k,n,m}^a\right)_{n,m\in\mathcal{N}_k}$ is the k^{th} company stationary one-period *state transition probability matrix* if action a is decided at the beginning of a period, i.e., $p_{k,n,m}^a$ is the probability of moving to state m from state n under action a; in particular, we have

$$\boldsymbol{P}_k^1 := \begin{array}{c} \\ * \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} * & 0 & 1 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & \mu_k & 1-\mu_k \end{pmatrix} \end{array}, \quad \boldsymbol{P}_k^0 := \begin{array}{c} \\ * \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} * & 0 & 1 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \theta_k & 1-\theta_k \end{pmatrix} \end{array}.$$

The state transition can be illustrated by the following schemes.

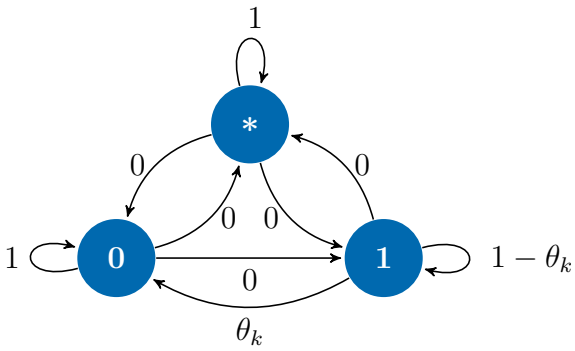

Figure 2.1: State transition of model 1 for action 0

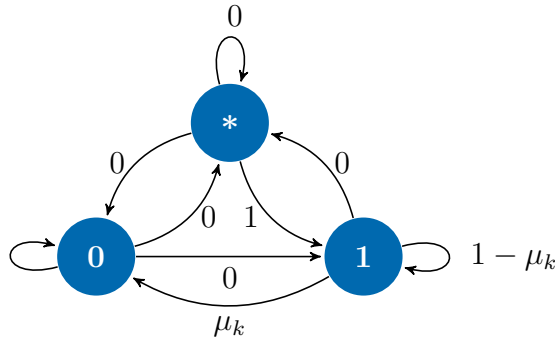

Figure 2.2: State transition of model 1 for action 1

The dynamics of company k is thus captured by action process $a_k(\cdot)$ and the state process $X_k(\cdot)$, which correspond to the actions $a_k(t) \in \mathcal{A}$ at all time instants $t \in \mathcal{T}$. As a result of deciding action $a_k(t)$ in state $X_k(t)$ at a time instant t, the company k provide the rewards for the investor, and evolves its state for the time instant $t+1$. At every state we have the same action space A available, which assures a technically useful property that $\boldsymbol{W}_k^a, \boldsymbol{R}_k^a$ and $\boldsymbol{P}_k^a$ are defined in the same dimensions under any $a \in \mathcal{A}$.

2.1.3 Multi-armed bandit problem and solution approach

In this section we introduce the Whittle relaxation for the general version of the multi-armed restless bandits. It turns out that the relaxation in fact allows to decompose the problem, and the optimal solution to the relaxed problem can be obtained by solving the single company subproblems. In the section 2.1.7 we show how the solution of the relaxed problem can be used to construct a heuristics for the original problem (P1) in finance.

General optimization problem

Whittle proposed his relaxation for the general case of the multi-armed restless bandits (see Jacko (2009b)). Thus, let us denote by $\mathbb{E}_{\mathbf{n}}^{\pi}$ the expectation conditioned on the joint initial state $\mathbf{n} := (n_k)_{k\in\mathcal{K}}$, where $X_k(0) = n_k$. For any initial joint state $\mathbf{n} = (n_k)_{k\in\mathcal{K}}$ and for any discount factor $0 < \beta < 1$, the discounted problem is to find an admissible policy $\pi \in \Pi$ maximizing the objective given by the expected discounted total reward, i.e.,

$$\max_{\pi} \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{k \in \mathcal{K}} \sum_{t=0}^{\infty} \beta^t R^{a_k(t)}_{k, X_k(t)} \right], \tag{P}$$

subject to the sample path capacity constraint,

$$\sum_{k \in \mathcal{K}} W^{a_k(t)}_{k, X_k(t)} \leq W \text{ , for all } t = 0, 1, 2, \ldots \tag{2.1}$$

where W is the available capacity to be used in every period. Since

$$\sum_{k \in \mathcal{K}} R^{a_k}_{k, X_k} \leq K \cdot \max_{a_k, X_k} R^{a_k}_{k, X_k}$$

the infinite series in (P) converge for $0 \leq \beta < 1$.

Analogously in the time-average criterion formulation instead of (P) we have

$$\max_{\pi} \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{k \in \mathcal{K}} \sum_{t=0}^{T-1} R^{a_k(t)}_{k, X_k(t)} \right], \tag{2.2}$$

subject to the same sample path capacity constraint (2.1).

Relaxations and decomposition

For large values of K the problem is analytically intractable, and therefore we approach it in different way. The main idea is to solve a modification of the problem (P) called Whittle's relaxation.

General Whittle and Lagrangian relaxation

Whittle (1988) proposed to relax the sample path capacity constraint (2.1), so under the discounted criterion we require this constraint to hold only in "expected total discounted " terms (Jacko (2009b)). We called it *Whittle relaxation*,

$$\mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{k \in \mathcal{K}} \sum_{t=0}^{\infty} \beta^t W^{a_k(t)}_{k, X_k(t)} \right] \leq \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{t=0}^{\infty} \beta^t W \right] = \frac{W}{1 - \beta} \tag{2.3}$$

and under the time-average criterion we require the constraint (2.1) to hold only in "expected time-average" terms,

$$\lim_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{k \in \mathcal{K}} \sum_{t=0}^{T-1} W^{a_k(t)}_{k, X_k(t)} \right] \leq \lim_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{t=0}^{T-1} W \right] = W. \tag{2.4}$$

The problems (P) and (2.2) with the relaxed constraints (2.3) and (2.4) respectively can be solved by Lagrangian relaxation (see, e.g., Guignard (2003) and Visweswaran (2009)), where we introduce a non-negative Lagrangian multiplier ν, to dualize the constraint (Jacko (2009b)). That is for the discounted criterion we obtain

$$\max_{\pi} \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{k \in \mathcal{K}} \sum_{t=0}^{\infty} \beta^t \left(R^{a_k(t)}_{k,X_k(t)} - \nu W^{a_k(t)}_{k,X_k(t)} \right) \right] + \nu \frac{W}{1-\beta}$$

and under the time-average criterion

$$\max_{\pi} \lim_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\mathbf{n}} \left[\sum_{k \in \mathcal{K}} \sum_{t=0}^{T-1} \left(R^{a_k(t)}_{k,X_k(t)} - \nu W^{a_k(t)}_{k,X_k(t)} \right) \right] + \nu W.$$

Above mentioned methods are applied on our model in the following sections.

A unified optimization criterion

Before describing the problem for model 1 we first define an averaging operator that allows us to discuss the infinite-horizon problem under the traditional discounted criterion and the time-average criterion in parallel. Let $\Pi_{X,a}$ be the set of all the policies that at each time instant t decide action $a(t)$ based only on the current state $X(t)$[1]. Let $\mathbb{E}^{\pi}_{\tau}$ denote the expectation over the state process $X(\cdot)$ and over the action process $a(\cdot)$, conditioned on the $X(\tau)$ and on policy π.

In general, consider any expected one-period variable $Q^{a(t)}_{X(t)}$ that depends on state $X(t)$ and on action $a(t)$ at any time instant t. For any policy $\pi \in \Pi_{X,a}$, any initial time instant $\tau \in \mathcal{T}$, and any *discount factor* $0 \leq \beta \leq 1$ we define the infinite-horizon *β-average quantity* of $Q^{a(t)}_{X(t)}$ as[2]

$$\mathbb{B}^{\pi}_{\tau} \left[Q^{a(\cdot)}_{X(\cdot)}, \beta, \infty \right] := \lim_{T \to \infty} \frac{\sum_{t=\tau}^{T-1} \beta^{t-\tau} \, \mathbb{E}^{\pi}_{\tau} \left[Q^{a(t)}_{X(t)} \right]}{\sum_{t=\tau}^{T-1} \beta^{t-\tau}}. \tag{2.5}$$

The β-average quantity recovers the traditionally considered quantities in the following three cases:

- *expected time-average quantity* when $\beta = 1$.
- *expected total discounted quantity*, scaled by constant $1 - \beta$, when $0 < \beta < 1$;

[1]Note that $X(t)$ and $a(t)$ include the information about the state-process history $X(0), X(1), \ldots, X(t-1)$ and about the action-process history $a(0), a(1), \ldots, a(t-1)$, due to Markov property

[2]For definiteness, we consider $\beta^0 = 1$ for $\beta = 0$.

- *myopic quantity* when $\beta = 0$.

Thus, when $\beta = 1$, the problem is formulated under the *time-average criterion*, whereas when $0 < \beta < 1$ the problem is considered under the *discounted criterion*. The remaining case when $\beta = 0$ reduces to a static problem and hence is considered in order to define a *myopic policy*. In the following we consider the discount factor β to be fixed and the horizon to be infinite, therefore we omit them in the notation and write briefly $\mathbb{B}^{\pi}_{\tau}\left[Q^{a(\cdot)}_{X(\cdot)}\right]$.

2.1.4 Optimization problem, relaxation and decomposition model 1

Optimization problem model 1

Based on the section (2.1.3) we now describe in more detail the problem we consider in model 1. Let $\Pi_{\boldsymbol{X},\boldsymbol{a}}$ be the space of non-anticipative policies depending on the joint state-process $\boldsymbol{X}(\cdot) := (X_k(\cdot))_{k\in\mathcal{K}}$ and deciding the joint action-process $\boldsymbol{a}(\cdot) := (a_k(\cdot))_{k\in\mathcal{K}}$, i.e., $\Pi_{\boldsymbol{X},\boldsymbol{a}}$ is the *joint policy space*. Further we denote $\mathbb{E}^{\pi}_{t}$ the expectation over the joint state process $\boldsymbol{X}(\cdot)$ and over the joint action process $\boldsymbol{a}(\cdot)$.

For any discount factor β, the problem is to find a joint policy $\boldsymbol{\pi}$ maximizing the *objective* given by the β-average aggregate reward starting from the initial time instant 0 subject to the family of *sample path* action constraints, i.e.

$$\begin{aligned} &\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}^{\boldsymbol{\pi}}_{0}\left[\sum_{k\in\mathcal{K}} R^{a_k(\cdot)}_{k,X_k(\cdot)}\right] \\ &\text{subject to } \left[\sum_{k\in\mathcal{K}} a_k(t)\right] = 1, \text{ for all } t\in\mathcal{T}. \end{aligned} \tag{P1}$$

Relaxations model 1

We use the fact that $W^{a_k(t)}_{k,X_k(t)} = a_k(t)$ (cf. definitions in 2.1.2) and instead of the constraints in (P1) we consider the sample path *consumption* constraints $\mathbb{E}^{\boldsymbol{\pi}}_{\tau}\left[\sum_{k\in\mathcal{K}} W^{a_k(t)}_{k,X_k(t)}\right] = 1$, for all $\tau \le t \in \mathcal{T}$. For $\tau = 0$ we obtain

$$\mathbb{E}^{\boldsymbol{\pi}}_{0}\left[\sum_{k\in\mathcal{K}} W^{a_k(t)}_{k,X_k(t)}\right] = 1, \text{ for all } t\in\mathcal{T} \tag{2.6}$$

requiring that the expected attention be fully allocated at every time instant if conditioned on $\boldsymbol{X}(0)$ only. Finally, we may require this constraint to hold only on β-average, as the *β-average capacity consumption constraint*

$$\mathbb{B}^{\boldsymbol{\pi}}_{0}\left[\sum_{k\in\mathcal{K}} W^{a_k(\cdot)}_{k,X_k(\cdot)}\right] = \mathbb{B}^{\boldsymbol{\pi}}_{0}\left[1\right]. \tag{2.7}$$

Using $\mathbb{B}_0^{\boldsymbol{\pi}}[1] = 1$, we obtain the following *Whittle relaxation* of problem (P1),

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}_0^{\boldsymbol{\pi}}\left[\sum_{k\in\mathcal{K}} R_{k,X_k(\cdot)}^{a_k(\cdot)}\right] \tag{P$^{\mathrm{W}}$}$$
$$\text{subject to } \mathbb{B}_0^{\boldsymbol{\pi}}\left[\sum_{k\in\mathcal{K}} W_{k,X_k(\cdot)}^{a_k(\cdot)}\right] = 1.$$

The *Whittle relaxation* (P$^{\mathrm{W}}$) can be treated by traditional Lagrangian methods, introducing a Lagrangian parameter, say ν, to dualize the constraint, obtaining thus the following Lagrangian relaxation,

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}_0^{\boldsymbol{\pi}}\left[\sum_{k\in\mathcal{K}} R_{k,X_k(\cdot)}^{a_k(\cdot)} - \nu\sum_{k\in\mathcal{K}} W_{k,X_k(\cdot)}^{a_k(\cdot)}\right] + \nu. \tag{P$_\nu^{\mathrm{L}}$}$$

The classic Lagrangian result (Guignard (2003),Visweswaran (2009)) says the following:

Proposition 2.1.1. *For every ν, P_ν^L provides an upper bound for the optimal value of both problem P^W and problem* (P1)*.*

Proof. A proof can be find in Niño-Mora (2001), for instance. □

Decomposition into single-company subproblems

We now decompose the optimization problem (P$_\nu^{\mathrm{L}}$) into isolated problems for each individual k, as it is standard for Lagrangian relaxations, considering ν as a parameter. That is, one can decide the action $a_k(t)$ independently of $X_j(t), j \neq k$. Notice that any joint policy $\boldsymbol{\pi} \in \Pi_{\boldsymbol{X},\boldsymbol{a}}$ defines a set of single-company policies $\widetilde{\pi}_k$ for all $k \in \mathcal{K}$, where $\widetilde{\pi}_k$ is a non-anticipative policy deciding the *company* k action-process $a_k(\cdot)$ depending on the *joint* state-process $\boldsymbol{X}(\cdot)$. We write $\widetilde{\pi}_k \in \Pi_{\boldsymbol{X},a_k}$. We therefore study the company k subproblem

$$\max_{\widetilde{\pi}_k\in\Pi_{\boldsymbol{X},a_k}} \mathbb{B}_0^{\widetilde{\pi}_k}\left[R_{k,X_k(\cdot)}^{a_k(\cdot)} - \nu W_{k,X_k(\cdot)}^{a_k(\cdot)}\right]. \tag{2.8}$$

2.1.5 Solution

In some cases, the problem (2.8) can be solved by assigning a set of index values $\nu_{k,n}$ to each state $n \in \mathcal{N}_k$. We refer to such cases as *indexable*. In the following, based on Jacko (2011), we characterize the index values $\nu_{k,n}$ to each state $n \in \mathcal{N}_k$. To do so we analytically emulate the AG-algorithm (see appendix (A. 3)), what is one of the main theoretical results of this thesis.

Optimal solution to single-company subproblem

As we mentioned above, the problem has to be indexable if we want to assign a set of index values which solve the problem. Therefore, we first define indexability similarly as in Jacko (2010b). Another definition of indexability can be found in the appendix (A. 1).

Definition 2.1.2. ***(Indexability)*** *We say that the problem* (2.8) *is indexable, if there exist values* $-\infty \leq \nu_{k,n} \leq \infty$ *for all* $n \in \mathcal{N}_k$ *such that the following holds for every state* $n \in \mathcal{N}_k$*:*

i) if $\nu \leq \nu_{k,1}$*, then* $a_k = 1$*, i.e. it is optimal to actively invest in the company* k *in state* 1*.*

ii) if $\nu > \nu_{k,1}$*, then* $a_k = 0$*, i.e. it is optimal not to actively invest in the company* k *in state* 1*.*

iii) if $\nu \leq \nu_{k,*}$*, then* $a_k = 1$*, i.e. it is optimal to actively invest in the company* k *in state* $*$*.*

iv) if $\nu > \nu_{k,*}$*, then* $a_k = 0$*, i.e. it is optimal not to actively invest in the company* k *in state* $*$*.*

v) if $\nu \leq \nu_{k,0}$*, then* $a_k = 1$*, i.e. it is optimal to actively invest in the company* k *in state* 0*.*

vi) if $\nu > \nu_{k,0}$*, then* $a_k = 0$*, i.e. it is optimal not to actively invest in the company* k *in state* 0*.*

The function $n \to \nu_{k,n}$ is called *(Whittle) index*, and $\nu_{k,n}$ are called the *(Whittle) index values*. Note that this definition is a generalization of the definitions introduced in Whittle (1988) and in Niño-Mora (2007), because we allow index values to be also equal to $-\infty$ and ∞. We are now ready to characterize the index values in closed form.

Theorem 2.1.3. *Suppose that the problem* (2.8) *is indexable, then the Whittle index values for the problem* (2.8) *are as follows,*

- *index value for a company* k *in the state* 0 *is*

$$\nu_{k,0} = 0,$$

- *index value for a company* k *in the state* $*$ *is*

$$\nu_{k,*} = \frac{-c_k^1 + \beta \mathcal{R}_k \mu_k}{(1 + \beta \mu_k)},$$

- *index value for a company* k *in the state* 1 *is*

$$\begin{aligned} \nu_{k,1} = & - c_k^1(1 - \mu_k) + c_k^0(1 - \theta_k) + \\ & + \beta c_k^1(1 - \mu_k - \theta_k + \theta_k \mu_k) - \beta c_k^0(1 - \mu_k - \theta_k + \theta_k \mu_k) + \\ & + \mathcal{R}_k \mu_k(1 - \beta + \beta \theta_k) + d_k \theta_k(1 - \beta + \beta \mu_k), \end{aligned}$$

so satisfy the indexability conditions i) - vi) stated in the definition (2.1.2)*.*

2.1.6 Proof of the theorem (2.1.3)

In the proof we analytically emulate the *AG-algorithm* proposed by Niño-Mora (2007). Its whole description and definition can be found in the appendix (A. 3). At this place we first recall some important concepts defined in appendix (A. 1) and then we briefly present main steps of this algorithm.

Important concepts from the appendix section (A. 1)

Let $\mathcal{S} \subset \{*, 0, 1\}$ be a subset of the set of states. Denote

$$a_k^{\mathcal{S}} = \chi_{\mathcal{S}}(X_k),$$

where

$$\chi_{\mathcal{S}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{S} \\ 0 & \text{if } x \notin \mathcal{S} \end{cases}$$

is the characteristic function of $\mathcal{S}$.

By choosing $\mathcal{S}$ we fully determine the decision rule for a_k for a particular k. We call $\mathcal{S}$ the *active set*.

The ν*-wage problem* (see appendix (A.3)) for the problem (2.8) is

$$\max_{\mathcal{S} \in \mathcal{N}_k} \mathbb{B}_0^{\mathcal{S}} \left[R_{k,X_k(\cdot)}^{a_k(\cdot)} \right] - \nu \, \mathbb{B}_0^{\mathcal{S}} \left[W_{k,X_k(\cdot)}^{a_k(\cdot)} \right]. \tag{2.9}$$

AG-algorithm basic description

As we present in the appendix (A. 2), the optimal policies to (2.9) lie on the upper boundary of the work-reward region and the parameter ν gives the slope of the supporting hyperplane defining an optimum point. Based on Niño-Mora (2007) we can find candidates $\nu_{k,n}^{\mathcal{S}}$ for the index value of company k, if the active set $\mathcal{S}$ is assumed. In other words $\nu_{k,n}^{\mathcal{S}}$ are indices $\nu_{k,n}$ for the particular company k, calculated under assumption that in the future an action is determined by the active set $\mathcal{S}$. We assume that the investor does not invest ($\mathcal{S}_0 = \emptyset$) and under this assumption we compute values $\nu_{k,n}^{\emptyset}$ for each state n. Then we arrange index value candidates into non-decreasing sequence and the largest candidate is proposed to be an index for that particular state n. This state is then incorporated into the active set. Thus in the next step l when we compute new candidates, we assume that we do not invest except for states which are already in the active set $\mathcal{S}_l$ and for the states which are not in the active set $\mathcal{S}_l$ we compute values $\nu_{k,n}^{\mathcal{S}_l}$ (we denote number of such states as i). The procedure is same as before. We repeat the same technique until all the states are in the active set ($\mathcal{S}_{\mathcal{N}} = \mathcal{N}_k$).

After adoption of the notation from the appendix (A. 3) the *adaptive-greedy algorithm* is

Algorithm *AG*
output: $\{n_l, \nu^*_{n_l}\}_{l=1}^{i}$
$\mathcal{S}_0 := \emptyset$
for $l := 1$ *to* i **do**
  pick $n_l \in \arg\max\left\{\nu_n^{\mathcal{S}_{l-1}} : n \in \partial^{out}_{\mathcal{F}}\mathcal{S}_{l-1}\right\}$;
  $\nu^*_{n_l} := \nu_{n_l}^{\mathcal{S}_{l-1}}; \mathcal{S}_l := \mathcal{S}_{l-1} \cup \{n_l\}$;
end

Algorithm 1: $AG - algorithm$

So we proceed in the way just described. Let us denote the optimal value function by $\widehat{V}_{k,n}$ for company k in state n. Then the Bellman equation for company k in states $n_k = \{0, *, 1\}$ respectively are

$$\widehat{V}_{k,0} = \max\{R^1_{k,0} - \nu W^1_{k,0} + \beta\widehat{V}_{k,0}; R^0_{k,0} - \nu W^0_{k,0} + \beta\widehat{V}_{k,0}\},$$

$$\widehat{V}_{k,*} = \max\{R^1_{k,*} - \nu W^1_{k,*} + \beta\widehat{V}_{k,1}; R^0_{k,*} - \nu W^0_{k,*} + \beta\widehat{V}_{k,*}\},$$

$$\widehat{V}_{k,1} = \max\{R^1_{k,1} - \nu W^1_{k,1} + \beta[\mu_k\widehat{V}_{k,0} + (1-\mu_k)\widehat{V}_{k,1}]; R^0_{k,1} - \nu W^0_{k,1} + \beta[\theta_k\widehat{V}_{k,0} + (1-\theta_k)\widehat{V}_{k,1}]\},$$

after substitution the formulas for expected rewards and expected one-period attention consumption, the Bellman equations are

$$\widehat{V}_{k,0} = \beta\widehat{V}_{k,0} + \max\{-\nu; 0\}, \tag{2.10}$$

$$\widehat{V}_{k,*} = \max\{-c^1_k - \nu + \beta\widehat{V}_{k,1}; \beta\widehat{V}_{k,*}\}, \tag{2.11}$$

$$\widehat{V}_{k,1} = \beta\widehat{V}_{k,1} + \max\{-c^1_k(1-\mu_k) + \mathcal{R}_k\mu_k - \nu - \beta\mu_k[\widehat{V}_{k,0} - \widehat{V}_{k,1}]; -c^0_k(1-\theta_k) - d_k\theta_k + \beta\theta_k[\widehat{V}_{k,0} - \widehat{V}_{k,1}]\}. \tag{2.12}$$

In all the Bellman equations the first term in braces correspond to active investing and the second to not active investing. We stated them before the first step of algorithm, because we use them in all the following steps for deriving balance equations $V^{\mathcal{S}}_{k,n}$ under policy $\mathcal{S}$. Now we can proceed to the first step of the algorithm.

1. **Step**. In the first step $\mathcal{S}_0 = \{\emptyset\}$, i.e. the investor does not invest actively in any of states. Recall that index value candidates are such values of $\nu_{k,n}^{\mathcal{S}_0}$ under the policy $\mathcal{S}_0$ that it does not matter if the investor invests or not. By deriving the balance equations for each state n_k we can prove the following lemma.

   Lema 2.1.4. *Under policy $\mathcal{S}_0 = \{\emptyset\}$, index value candidates are*

$$\nu_{k,0}^{\mathcal{S}_0} = 0$$

$$\nu_{k,*}^{\mathcal{S}_0} = -c_k^1 + \beta\left(\frac{-c_k^0(1+\theta_k) - d_k\theta_k}{1-\beta+\beta\theta_k}\right)$$

$$\begin{aligned}\nu_{k,1}^{\mathcal{S}_0} = &- c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \\ &+ \beta c_k^1(1-\mu_k-\theta_k+\theta_k\mu_k) - \beta c_k^0(1-\mu_k-\theta_k+\theta_k\mu_k) + \\ &+ \mathcal{R}_k\mu_k(1-\beta+\beta\theta_k) + d_k\theta_k(1-\beta+\beta\mu_k).\end{aligned}$$

   Proof. The equation (2.10) leads us to the index value candidate $\nu_{k,0}^{\mathcal{S}_0} = 0$ corresponding to state 0. In the other words, if $\nu = \nu_{k,0}^{\mathcal{S}_0}$ in (2.10) it does not matter if the investor invests or not. Because in step 1 the active set is $\mathcal{S}_0 = \{\emptyset\}$, so the investor does not invest actively in any of states, we can write the balance equation for state 0

$$V_{k,0}^{\mathcal{S}_0} = \beta V_{k,0}^{\mathcal{S}_0} + 0.$$

   Thus, $V_{k,0}^{\mathcal{S}_0} = 0$ and under $\mathcal{S}_0 = \{\emptyset\}$ it has to hold that $\nu \geq \nu_{k,0}^{\mathcal{S}}$, what can be observed from equation(2.10).

   Based on the equation (2.11) we derive the balance equation for state $*$

$$V_{k,*}^{\mathcal{S}_0} = \beta V_{k,*}^{\mathcal{S}_0},$$

   what lead us to $V_{k,*}^{\mathcal{S}_0} = 0$. Moreover, from the equation (2.11) we get the condition for index value candidate $\nu_{k,*}^{\mathcal{S}_0}$ that must be satisfied under $\mathcal{S}_0 = \{\emptyset\}$

$$\nu_{k,*}^{\mathcal{S}_0} = -c_k^1 + \beta(V_{k,1}^{\mathcal{S}_0} - V_{k,*}^{\mathcal{S}_0}) = -c_k^1 + \beta V_{k,1}^{\mathcal{S}_0}. \tag{2.13}$$

   To continue we now need to compute $V_{k,1}^{\mathcal{S}_0}$. Similarly as before, based on the equation (2.12) and on the $\mathcal{S}_0 = \{\emptyset\}$ we obtain

$$V_{k,1}^{\mathcal{S}_0} = \frac{-c_k^0(1-\theta_k) - d_k\theta_k}{(1-\beta+\beta\theta_k)}. \tag{2.14}$$

By substituting $V_{k,1}^{\mathcal{S}_0}$ into the condition (2.13) we get the index value candidate corresponding to state $*$

$$\nu_{k,*}^{\mathcal{S}_0} = -c_k^1 + \beta\left(\frac{-c_k^0(1-\theta_k) - d_k\theta_k}{1-\beta+\beta\theta_k}\right).$$

For obtaining the last index value candidate corresponding to state 1 we derive the index value candidate $\nu_{k,1}^{\mathcal{S}_0}$ from equation (2.12)

$$\nu_{k,1}^{\mathcal{S}_0} = -c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k + \beta V_{k,1}^{\mathcal{S}_0}(\theta_k - \mu_k),$$

and by substitution for $V_{k,1}^{\mathcal{S}_0}$ we have

$$\begin{aligned}\nu_{k,1}^{\mathcal{S}_0} =& -c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \\ &+ \beta c_k^1(1-\mu_k-\theta_k+\theta_k\mu_k) - \beta c_k^0(1-\mu_k-\theta_k+\theta_k\mu_k) + \\ &+ \mathcal{R}_k\mu_k(1-\beta+\beta\theta_k) + d_k\theta_k(1-\beta+\beta\mu_k).\end{aligned}$$

□

Note that for each state holds $\nu \geq \nu_{k,0}^{\mathcal{S}_0}$, $\nu \geq \nu_{k,*}^{\mathcal{S}_0}$ and $\nu \geq \nu_{k,1}^{\mathcal{S}_0}$, so that the investor does not invest actively in any of them.

For identification which of the index value candidates is the index value, we need to find the largest index value candidate.

Lema 2.1.5. *Under policy $\mathcal{S}_0 = \{\emptyset\}$ holds that $\nu_{k,1}^{\mathcal{S}_0} \geq \nu_{k,0}^{\mathcal{S}_0} \geq \nu_{k,*}^{\mathcal{S}_0}$.*

Proof. Let start from the end and compare $\nu_{k,0}^{\mathcal{S}_0}$ with $\nu_{k,*}^{\mathcal{S}_0}$. Because $0 < \beta < 1$ we can see that all three terms in $\nu_{k,*}^{\mathcal{S}_0}$ are negative, what lead us to $\nu_{k,0}^{\mathcal{S}_0} \geq \nu_{k,*}^{\mathcal{S}_0}$ because $\nu_{k,0}^{\mathcal{S}_0} = 0$.

Now we compare larger index value candidate from the previous comparison with the last one, i.e. $\nu_{k,1}^{\mathcal{S}_0}$ with $\nu_{k,0}^{\mathcal{S}_0}$. Recall that

$$\begin{aligned}\nu_{k,1}^{\mathcal{S}_0} =& -c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \\ &+ \beta c_k^1(1-\mu_k-\theta_k+\theta_k\mu_k) - \beta c_k^0(1-\mu_k-\theta_k+\theta_k\mu_k) + \\ &+ \mathcal{R}_k\mu_k(1-\beta+\beta\theta_k) + d_k\theta_k(1-\beta+\beta\mu_k).\end{aligned}$$

We can easily rearrange it to

$$(1-\beta+\beta\theta_k)(c_k^1\mu_k - c_k^1 + \mathcal{R}_k\mu_k) + (1-\beta+\beta\mu_k)(c_k^0 - c_k^0\theta_k + d_k\theta_k).$$

All parentheses containing β are bigger than 0, because $0 < \beta < 1$. The term $c_k^0 - c_k^0\theta_k + d_k\theta_k = c_k^0(1-\theta_k) + d_k\theta_k$ is also bigger or equal to zero. The reason is that all parameters are not negative and $0 \leq \theta \leq 1$. The last term $c_k^1\mu_k - c_k^1 + \mathcal{R}_k\mu_k$ is non-negative under condition $\mathcal{R}_k \geq \frac{c_k^1}{\mu_k} - c_k^1$, which hold by assumption, because $\mathcal{R}_k \geq \frac{c_k^1}{\beta\mu_k}$ always implies $\mathcal{R}_k \geq \frac{c_k^1}{\mu_k} - c_k^1$. We just showed that $\nu_{k,1}^{\mathcal{S}_0} \geq 0$, so $\nu_{k,1}^{\mathcal{S}_0} \geq \nu_{k,0}^{\mathcal{S}_0}$, because $\nu_{k,0}^{\mathcal{S}_0} = 0$. □

We can finally write $\nu_{k,1}^{\mathcal{S}_0} \geq \nu_{k,0}^{\mathcal{S}_0} \geq \nu_{k,*}^{\mathcal{S}_0}$, in other words $\nu_{k,1}^{\mathcal{S}_0}$ is not only the candidate for the index value, but it is an index value $\nu_{k,1}$ for companies that are in state 1. Therefore, in the next step state 1 is included into the active set $\mathcal{S}_1$.

Note that in this step we proved that statements *i)* and *ii)* from definition (2.1.2) holds. These statements claim that it is optimal to actively invest if $\nu \leq \nu_{k,1}$ and it is optimal not to actively invest if $\nu \geq \nu_{k,1}$. This follows directly from the equation (2.12).

2. **Step**. In the second step we include state 1 into the active set $\mathcal{S}_1$, so state 1 is the only state in which the investor invests actively and in the other two states she does not invest actively, i.e. $\mathcal{S}_1 = \{1\}$. Similarly as in the step 1, using the balance equations based on the Bellman equations we can prove the following lemma.

Lema 2.1.6. *Index candidates for each state $n_k \notin \mathcal{S}_1$, where the active set $\mathcal{S}_1 = \{1\}$ are*

$$\nu_{k,0}^{\mathcal{S}_1} = 0$$

$$\nu_{k,*}^{\mathcal{S}_1} = \frac{-c_k^1 + \beta\mathcal{R}_k\mu_k}{(1+\beta\mu_k)}$$

Proof. In this step we can use previous results for states 0 and $*$. For state 0 the situation is totally the same as in the previous section, because it is not in $\mathcal{S}_1$. For state $*$ the situation is similar, but for computation of the index value candidate we need to substitute $V_{k,1}^{\mathcal{S}_1}$. This is different to the first step, because the investor invests actively in state 1. Specifically for state 0 we have a candidate for an index value $\nu_{k,0}^{\mathcal{S}_1} = 0$ what directly proves the first part of the lemma (2.1.6); and for state $*$ we have $V_{k,*}^{\mathcal{S}_1} = 0$ and that has to be satisfied $\nu_{k,*}^{\mathcal{S}_1} = -c_k^1 + \beta V_{k,1}^{\mathcal{S}_1}$. Therefore we need to derive $V_{k,1}^{\mathcal{S}_1}$ under assumption that the

investor invests actively in state 1. Straightforwardly from the equation (2.12) we can write

$$V_{k,1}^{\mathcal{S}_1} = \beta V_{k,1}^{\mathcal{S}_1} - c_k^1(1-\mu_k) + \mathcal{R}_k\mu_k - \nu + \beta\mu_k(V_{k,0}^{\mathcal{S}_1} - V_{k,1}^{\mathcal{S}_1}).$$

Rearranging of terms lead us to

$$V_{k,1}^{\mathcal{S}_1} = \frac{-c_k^1(1-\mu_k) + \mathcal{R}_k\mu_k - \nu}{(1-\beta+\beta\mu_k)}$$

Now we can return to state $*$ and continue with computation of the index value candidate for state $*$. After substitution for $V_{k,1}^{\mathcal{S}_1}$ into the equation for $\nu_{k,*}^{\mathcal{S}_1}$ that must be satisfied, we obtain the index value candidate for state $*$

$$\nu_{k,*}^{\mathcal{S}_1} = \frac{-c_k^1 + \beta\mathcal{R}_k\mu_k}{(1+\beta\mu_k)},$$

for which under $\mathcal{S}_1$ holds that $\nu \geq \nu_{k,*}^{\mathcal{S}}$. □

When we have the index value candidates for states $*$ and 0, we need to find which one of them is larger. That leads us to the following lemma.

Lema 2.1.7. *Index value candidates satisfy* $\nu_{k,*}^{\mathcal{S}_1} \geq \nu_{k,0}^{\mathcal{S}_1}$ *for the active set* $\mathcal{S}_1 = \{1\}$.

Proof. We want to compare two index value candidates $\nu_{k,*}^{\mathcal{S}_1}$ and $\nu_{k,0}^{\mathcal{S}_1}$. We can easily observe that $\nu_{k,*}^{\mathcal{S}_1}$ is always non-negative, because condition

$$\mathcal{R} \geq \frac{c_k^1}{\beta\mu_k},$$

is satisfied due to assumptions made in the model description (2.1.2). The index value candidate corresponding to state 0 equals to zero, so we proved the lemma (2.1.7). □

Therefore, we obtained an index value for state $\nu_{k,*} = \frac{-c_k^1+\beta\mathcal{R}_k\mu_k}{(1+\beta\mu_k)}$. The statements *iii)* and *iv)* from definition (2.1.2), which claim that it is optimal to actively invest in the company k in state $*$ if $\nu \leq \nu_{k,*}$ and that it is optimal not to actively invest in the company k in state $*$ if $\nu \geq \nu_{k,*}$, hold for this model.

3. **Step.** In the third and the last step our initial active set is $\mathcal{S}_2 = \{*, 1\}$, thus the investor does not invest actively only in the state 0 and it is the only state for which we do not have an index. Thus we have only one candidate for an index value, what directly leads us to the conclusion that this candidate is also an index value. For state 0 we obtain an index value:

$$\nu_{k,0} = 0$$

Therefore, after this step the active set is $\mathcal{S}_3 = \{*, 0, 1\}$ so we can end the algorithm. Similarly as in the previous cases the validity of statements *v)* and *vi)* is a consequence of the equation (2.10).

2.1.7 Optimal solution to relaxations model 1

The vector of policies $\boldsymbol{\pi}^* := (\widetilde{\pi}_k^*)_{k\in\mathcal{K}}$ identified in theorem (2.1.3) consists of mutually independent single-company optimal policies, therefore this vector is an optimal policy to the Lagrangian relaxation $(\mathrm{P}_\nu^{\mathrm{L}})$.

Since a finite-state MDP admits an LP formulation using the standard *state-action frequency* variables (as observed in Niño-Mora (2001)), strong LP duality implies that there exists $\widehat{\nu}$ (possibly depending on the joint initial state) such that the Lagrangian relaxation $(\mathrm{P}_{\widehat{\nu}}^{\mathrm{L}})$ achieves the same objective value as $(\mathrm{P}^{\mathrm{W}})$. Further, if $\widehat{\nu} \neq 0$, then LP complementary slackness ensures that the β-average capacity constraint (2.7) is satisfied by any optimal solution to $(\mathrm{P}_{\widehat{\nu}}^{L})$. (see Jacko (2010a))

Index rule for the original problem model 1

Since the original problem requires to invest money in exactly one company, then at any time instant t we propose to invest money in the company $\widehat{k}(t)$ with the highest actual index, which correspond to the shadow price of investing in the company $\widehat{k}(t)$, i.e.,

$$\widehat{k}(t) := \arg\max_{k\in\mathcal{K}} \nu_{k,X_k(t)}.$$

Note that the investor invests in the company with the highest work efficiency. In other words, the investor chooses a company where it would be worth to work even if she had to pay for it the most.

Let recall the indices. For $0 < \beta < 1$, the index value for company k is one of the following three depending on the state of company k

$$\nu_{k,0} = 0,$$

$$\nu_{k,*} = \frac{-c_k^1 + \beta\mathcal{R}_k\mu_k}{(1+\beta\mu_k)},$$

$$\begin{aligned}\nu_{k,1} = &- c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \\ &+ \beta c_k^1(1-\mu_k-\theta_k+\theta_k\mu_k) - \beta c_k^0(1-\mu_k-\theta_k+\theta_k\mu_k) + \\ &+ \mathcal{R}_k\mu_k(1-\beta+\beta\theta_k) + d_k\theta_k(1-\beta+\beta\mu_k).\end{aligned}$$

Under $\beta = 1$, we obtain the time-average version of the index values

$$\bar{\nu}_{k,0} = 0,$$

$$\bar{\nu}_{k,*} = \frac{-c_k^1 + \mathcal{R}_k\mu_k}{(1+\mu_k)},$$

$$\bar{\nu}_{k,1} = c_k^1(-\theta_k+\theta\mu_k) + c_k^0(\mu_k-\theta\mu_k) + \mathcal{R}_k\mu_k\theta_k + d_k\theta_k\mu_k.$$

Finally, we just remark that $\beta = 0$ gives rise to the myopic version of the index values

$$\tilde{\nu}_{k,0} = 0,$$

$$\tilde{\nu}_{k,*} = -c_k^1,$$

$$\tilde{\nu}_{k,1} = -c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k.$$

2.2 Venture capitalists investments model 2

In the following two sections we only point out the differences to model 1. Everything not stated in these sections is the same as in model 1, if not stated otherwise.

2.2.1 Problem description model 2

In this model the investor can choose M companies where she invests. The difference from the previous model is that she must continuously be actively investing in order to keep a company in the portfolio. So, the two actions are "include in the portfolio" and "exclude from the portfolio". Thus, if the company succeeds in that time period investor does not earn anything from it. In other words, she would have net loss. When she stops to invest in the company, she can prevent to hold a bankrupted company. Therefore, we no longer make difference between active and passive investments, i.e. we have only one cost c_k. Another difference from the first model is that

the penalty cost for the company's bankruptcy is not the same as in the model 1. The penalty cost is d_k^1 if the investor invested in the company before its bankruptcy and d_k^0 is the penalty cost if the company succeeds but the investor did not invest in it before. The idea behind d_k^0 is that public opinion about the investor is getting worse, because she has wasted the opportunity to invest in the successful company. The investor's action is either to invest or not to invest. This model allows us to simulate budget constraints by setting the number of possible investments at one time period.

To simulate the real world better, we assume that the investor invests simultaneously to M companies on average ($W = M$ in the general case (see equation (2.1))) in one time period. Therefore, when the market is in a good condition, the investor can invest in more than to M companies, and to less than M companies if the market is in the opposite situation. It was proven by Whittle (1988) that the index policy is optimal, if the constraint is that the investor invests to M companies only on average. The rest of the investment characterization is the same as in the previous model.

2.2.2 MDP formulation model 2

Companies and industries model 2

The company k definition differs from the definition in model 1 only in the following parameters

- $\mathcal{N}_k := \{0, 1\}$ is the *state space*, where 0 represents a company that has bankrupted or succeeded; and state 1 means that the company is available for the investment, i.e. it is still in the market.
- $\boldsymbol{R}_k^a := \left(R_{k,n}^a\right)_{n \in \mathcal{N}_k}$, where $R_{k,n}^a$ is the expected one-period *reward* earned by the investor for company k at state n if action a is decided at the beginning of a period; in particular,

$$\begin{aligned} R_{k,0}^1 &:= 0, & R_{k,1}^1 &:= -c_k(1 - \mu_k - \theta_k) + R\mu_k - d_k^1\theta_k, \\ R_{k,0}^0 &:= 0, & R_{k,1}^0 &:= 0 \cdot \theta_k - d_k^0\mu_k \end{aligned}$$

  Where $R > c_k$, $d \geq c_k$ and $d_k^1 > d_k^0$.

- $\boldsymbol{P}_k^a := \left(p_{k,n,m}^a\right)_{n,m \in \mathcal{N}_k}$ is the k^{th} company stationary one-period *state-transition probability matrix* if action a is decided at the beginning of a time period, i.e., $p_{k,n,m}^a$ is the probability of moving to state m from state n under action a; in particular, we have

$$\boldsymbol{P}_k^1 = \boldsymbol{P}_k^0 := \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1 & 0 \\ (\mu_k + \theta_k) & (1 - \mu_k - \theta_k) \end{pmatrix} \end{array}.$$

The state transition can be illustrated by the scheme in the figure (2.3).

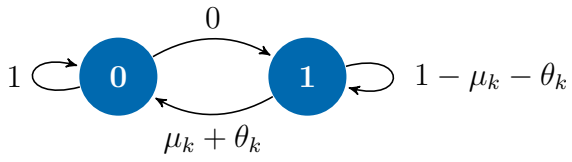

Figure 2.3: State transition of model 2 for action 1 and also for action 0

Optimization problem model 2

The optimization problem differs from the optimization problem (P1) for the model 1 in the constraint, which states that in every time instant we invest in M companies only on average. The problem, for any discount factor β, is to find a joint policy $\boldsymbol{\pi}$ maximizing the *objective* given by the β-average aggregate reward starting from the initial time instant 0 subject to the family of allocation constraints, i.e.,

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}_0^{\boldsymbol{\pi}}\left[\sum_{k\in\mathcal{K}} R_{k,X_k(\cdot)}^{a_k(\cdot)}\right] \tag{P2}$$
$$\text{subject to } \mathbb{B}_0^{\boldsymbol{\pi}}\left[\sum_{k\in\mathcal{K}} a_k(t)\right] = M.$$

In this model we do not need to relax the problem, because it is already in the form which we will obtain by the Whittle relaxation. The consequence of it is that we obtain the optimal solution for the original problem, not only a nearly-optimal. Now we can continue with decomposition into the single company subproblems. In the following we do not need to take care of the fact that we want to invest in M companies. The solution method is similar as before, the problem after Lagrangian relaxation and decomposition is the same as in the model 1 (formula (2.8)) , so we study the company k subproblem

$$\max_{\widetilde{\pi}_k\in\Pi_{\boldsymbol{X},a_k}} \mathbb{B}_0^{\widetilde{\pi}_k}\left[R_{k,X_k(\cdot)}^{a_k(\cdot)} - \nu W_{k,X_k(\cdot)}^{a_k(\cdot)}\right]. \tag{2.15}$$

2.2.3 Solution model 2

In this section we identify a set of optimal policies $\widetilde{\pi}_k^*$ to (2.15) for all companies $k\in\mathcal{K}$, and using them we construct a joint policy $\boldsymbol{\pi}$ optimal for the problem (P2).

Optimal solution to single-company subproblem model 2

Problem (2.15) falls into the framework of *restless bandits* and can be optimally solved by assigning a set of index values $\nu_{k,n}$ to each state $n\in\mathcal{N}_k$ under certain conditions

(Niño-Mora (2007)).

Let us denote for company k, $\nu_{k,0} := 0$ and

$$\nu_{k,1} := -c_k(1-\mu_k-\theta_k) + \mathcal{R}_k\mu_k - d_k^1\theta_k + d_k^0\mu_k$$

Then we can prove the following result.

Proposition 2.2.1. ***(Indexability)*** *For problem* (2.15) *and company k, the following holds:*

i) if $\nu \leq \nu_{k,1}$, then $a_k = 1$, i.e. it is optimal to invest in company k in state 1;

ii) if $\nu > \nu_{k,1}$, then $a_k = 0$, i.e. it is optimal not to invest in company k in state 1;

iii) if $\nu \leq \nu_{k,0}$, then $a_k = 1$, i.e. it is optimal to invest in company k in state 0;

iv) if $\nu > \nu_{k,0}$, then $a_k = 0$, i.e. it is optimal not to invest in company k in state 0;

Proof. Similarly to model 1 to prove this proposition we need to establish indexability of the problem and compute the index values following the survey Niño-Mora (2007). Indexability is in fact equivalent to existence of the quantities with stated properties, and is valid because any binary-state MDP is indexable (Niño-Mora (2007)). On the other hand, in disparity with the model 1 we do not need to use the AG-algorithm, because we have only two states and one of them corresponds to the company that has already succeeded or bankrupted. Therefore, we adopt the similar proof from Novak (2011) to this model.

Let us denote the optimal value function by $\widehat{V}_{k,n}$ for company k and state n. The Bellman equation for state 1 and company k, after substitution of definitions of the action-dependent parameters, the formulas for expected rewards and expected one-period attention consumption for a state 1, we obtain

$$\begin{aligned}\widehat{V}_{k,1} =& \beta\left[(\mu_k+\theta_k)\widehat{V}_{k,0} + (1-\mu_k-\theta_k)\widehat{V}_{k,1}\right] - d_k^0\mu_k + \\ &+ \max\{-c_k(1-\mu_k-\theta_k) + \mathcal{R}_k\mu_k - d_k^1\theta_k + d_k^0\mu_k - \nu; 0\}\end{aligned} \tag{2.16}$$

where the first term in the braces corresponds to investing into the company and the second term corresponds to not investing.

The Bellman equation for $\widehat{V}_{k,0}$ is:

$$\widehat{V}_{k,0} = \max\{0 - \nu + \beta\widehat{V}_{k,0}; \beta\widehat{V}_{k,0}\}. \tag{2.17}$$

From the Bellman equation (2.16) we can see that for deciding whether it is optimal to invest or not to invest in state 1, we do not need to know $\widehat{V}_{k,0}$ and $\widehat{V}_{k,1}$. If we want to invest the first term in braces should be larger than the second in the equation (2.16). We can write condition for investing to the company in state 1:

$$-c_k(1-\mu_k-\theta_k)+\mathcal{R}_k\mu_k-d_k^1+d_k^0\mu_k\geq\nu$$

We propose the term on left side of the inequality to be $\nu_{k,1}$, thus we proved the statement (2.2.1).

Analogous to above we can show that if

$$\nu_{k,1}\leq\nu,$$

then the action corresponding to not investing is larger than the action corresponding to investing. In other words, it is optimal not to invest (that proves the statement *ii)* in proposition (2.2.1)).

Similarly goes the proof for statements *iii)* and *iv)* in the proposition 2.2.1. Under assumption that the investor invests we can derive $\widehat{V}_{k,0}=\frac{\nu}{(1-\beta)}$ from the Bellman equation (2.17). In the case, that we assume that the investor does not invest, we derive that

$$\widehat{V}_{k,0}=0.$$

Therefore, when it is optimal to invest, action 1 is better than or equal to action 2, we obtain condition:

$$-\nu-\beta\left(\frac{\nu}{1-\beta}\right)\geq-\beta\left(\frac{\nu}{1-\beta}\right)$$

We have assumed that $\nu\leq\nu_{k,0}$ and we know that $\nu_{k,0}=0$. From the condition, we can observe that $\nu\leq 0$, so it is satisfied.

Analogous for the case, when it is optimal not to invest (action 1 is worse than or equal to action 2), we obtain similar condition that is satisfied when $\nu\geq 0$. Because we assumed that $\nu\geq\nu_{k,0}$, where $\nu_{k,0}=0$ it is also true.

□

Index rule for the problem model 2

The problem requires to invest money in M companies, then at any time instant t we propose to invest money to the M companies with the highest actual index $\nu_{k,X_k(t)}$, for which hold that $\nu_{k,X_k(t)}\geq 0$.

Note that in this case the time-average version of indices, myopic version and version under discounted criterion are the same. These indices for each state $n_k\in\mathcal{N}_k$ are

$$\nu_{k,0}=0$$

$$\nu_{k,1} = -c_k(1 - \mu_k - \theta_k) + \mathcal{R}_k\mu_k - d_k^1\theta_k + d_k^0\mu_k$$

It is also interesting to note that being myopic is optimal for the long-term, both discounted and time-average.

2.3 Venture capitalists investments model 3

2.3.1 Problem description model 3

Equivalently as in the model 2, we point out only the differences with the model 1. Model 3 takes the important features from both model 1 and model 2, now represents investments in industries rather than in companies. The principal modification from model 1 is that investments do not leave to state 0, but they come back to state $*$, so the investor can again invest in them. Contrary to model 2, passive investment does not exclude the company from the portfolio. This leads us to assumption that we look on companies from the industry level. Therefore, we assume that each industry is represented by a typical company for that particular industry.

2.3.2 MDP formulation model 3

Companies and industries model 3

The differences between model 1 and model 3 are only in the following parameters

- $\mathcal{N}_k := \{*, 1\}$ is the *state space*, where state $*$ represents a company from a particular industry without any investment or a company that had already been in investor's portfolio, but it either succeeded or bankrupted, and state 1 means that a company is in investors portfolio, (he had invested, but the company neither succeeded nor bankrupted);

- $\boldsymbol{R}_k^a := \left(R_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $R_{k,n}^a$ is the expected one-period *reward* earned by investor for company k at state n if action a is decided at the beginning of a period; in particular,

$$\begin{aligned} R_{k,*}^1 &:= -c_k^1, & R_{k,1}^1 &:= -c_k^1 \cdot (1 - \mu_k) + \mathcal{R}\mu_k, \\ R_{k,*}^0 &:= 0, & R_{k,1}^0 &:= -c_k^0 \cdot (1 - \theta_k) - d_k\theta_k \end{aligned}$$

  Where $\mathcal{R} > c_k^1 > c_k^0$ and $d > c_k^0$.

- $\boldsymbol{P}_k^a := \left(p_{k,n,m}^a\right)_{n,m\in\mathcal{N}_k}$ is the k^{th} company stationary one-period *state-transition probability matrix* if action a is decided at the beginning of a period, i.e., $p_{k,n,m}^a$ is the probability of moving to state m from state n under action a; in particular, we have

$$\boldsymbol{P}_k^1 := \begin{matrix} \\ * \\ 1 \end{matrix} \begin{matrix} \begin{matrix} * & 1 \end{matrix} \\ \begin{pmatrix} 0 & 1 \\ \mu_k & 1-\mu_k \end{pmatrix} \end{matrix}, \quad \boldsymbol{P}_k^0 := \begin{matrix} \\ * \\ 1 \end{matrix} \begin{matrix} \begin{matrix} * & 1 \end{matrix} \\ \begin{pmatrix} 1 & 0 \\ \theta_k & 1-\theta_k \end{pmatrix} \end{matrix}.$$

The state transition can be illustrated by the schemes in figures (2.4) and (2.5).

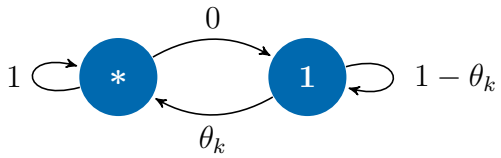

Figure 2.4: State transition of model 3 for action 0

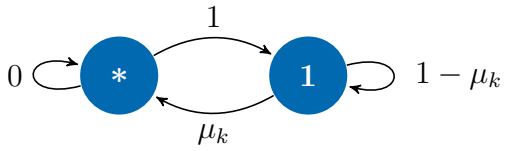

Figure 2.5: State transition of model 3 for action 1

Optimization problem VCs investment model 3

The optimization problem is the same as the optimization problem from the model 1, so it states that in every time instant we invest in exactly one company. The problem, for any discount factor β, is to find a joint policy $\boldsymbol{\pi}$ maximizing the *objective* given by the β-average aggregate reward starting from the initial time instant 0 subject to the family of *sample path* action constraints, i.e.,

$$\max_{\boldsymbol{\pi} \in \Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}_0^{\boldsymbol{\pi}} \left[\sum_{k \in \mathcal{K}} R_{k,X_k(\cdot)}^{a_k(\cdot)} \right] \tag{P3}$$

$$\text{subject to } \left[\sum_{k \in \mathcal{K}} a_k(t) \right] = 1, \text{ for all } t \in \mathcal{T}.$$

In this model we again need to relax the problem. Relaxation and decomposition is the same as in the model 1, so it leads us to the same company k subproblem as (2.8), so we study the company k subproblem

$$\max_{\widetilde{\pi}_k \in \Pi_{\boldsymbol{X},a_k}} \mathbb{B}_0^{\widetilde{\pi}_k} \left[R_{k,X_k(\cdot)}^{a_k(\cdot)} - \nu W_{k,X_k(\cdot)}^{a_k(\cdot)} \right]. \tag{2.18}$$

2.3.3 Solution model 3

Similarly as in the previous models we identify in this section a set of optimal policies $\widetilde{\pi}_k^*$ to (2.18) for all companies k, and we use them for constructing a joint policy $\boldsymbol{\pi}$ which is not necessarily optimal for problem (P3).

Optimal solution to single-company subproblem model 3

Problem (2.18) can be solved by assigning a set of index values $\nu_{k,n}$ to each state $n \in \mathcal{N}_k$ under condition of indexability (Niño-Mora (2007)), because it falls into the restless bandits framework. Therefore, we first prove the indexability for this problem.

Proposition 2.3.1. ***(Indexability)*** *The problem* (2.18) *is indexable and values* $-\infty \leq \nu_{k,n} \leq \infty$ *exists for all* $n \in \mathcal{N}_k$ *such that the following holds for every state* $n \in \mathcal{N}_k$*:*

i) if $\nu \leq \nu_{k,1}$*, then* $a_k = 1$*, i.e. it is optimal to actively invest in the company* k *in state* 1*.*

ii) if $\nu > \nu_{k,1}$*, then* $a_k = 0$*, i.e. it is optimal not to actively invest in the company* k *in state* 1*.*

iii) if $\nu \leq \nu_{k,*}$*, then* $a_k = 1$*, i.e. it is optimal to actively invest in the company* k *in state* $*$*.*

iv) if $\nu > \nu_{k,*}$*, then* $a_k = 0$*, i.e. it is optimal not to actively invest in the company* k *in state* $*$*.*

Proof. Indexability is valid because in Niño-Mora (2007) was proven that any binary-state MDP is indexable. □

Now we can characterize the index values for problem (2.18).

Theorem 2.3.2. *The Whittle index values for the problem* (2.18) *are as follows,*

- *index value for a company* k *in the state* $*$ *is*

$$\nu_{k,*} = \frac{-c_k^1(1-\beta+\beta\mu_k)}{(1+\beta\mu_k)} + \beta\left(\frac{-c_k^1(1-\mu_k)+\mathcal{R}_k\mu_k}{1+\beta\mu_k}\right) = \frac{-c_k^1+\beta\mathcal{R}_k\mu_k}{1+\beta\mu_k}$$

- *index value for a company* k *in the state* 1 *is*

$$\nu_{k,1} := -c_k^1(1-\mu_k) + \mathcal{R}\mu_k + \left[c_k^0(1-\theta_k) + d_k\theta_k\right] \cdot \left(\frac{1-\beta+\beta\mu_k}{1-\beta+\beta\theta_k}\right)$$

so satisfy the indexability conditions i) - iv) from the indexability proposition (2.3.1).

2.3.4 Proof of the theorem (2.3.2)

The proof goes similarly as in the model 1 and we use again the AG-algorithm, but first we introduce the Bellman equations for both states.

The Bellman equation for state $*$ and company k is

$$\widehat{V}_{k,*} = \max\{R_{k,*}^1 - \nu W_{k,*}^1 + \beta\widehat{V}_{k,1}; R_{k,*}^0 - \nu W_{k,*}^0 + \beta\widehat{V}_{k,*}\},$$

and for state 1 and company k is,

$$\widehat{V}_{k,1} = \max\{R^1_{k,1} - \nu W^1_{k,1} + \beta[\mu_k \widehat{V}_{k,*} + (1-\mu_k)\widehat{V}_{k,1}];$$
$$R^0_{k,1} - \nu W^0_{k,1} + \beta[\theta_k \widehat{V}_{k,*} + (1-\theta_k)\widehat{V}_{k,1}]\},$$

after substitution for expected rewards and expected one-period attention consumption, the Bellman equations for states $*$ and 1 respectively are

$$\widehat{V}_{k,*} = \max\{-c^1_k - \nu + \beta\widehat{V}_{k,1}; \beta\widehat{V}_{k,*}\}, \tag{2.19}$$

$$\begin{aligned}\widehat{V}_{k,1} = \max\{&-c^1_k(1-\mu_k) + \mathcal{R}_k\mu_k - \nu + \beta[\mu_k\widehat{V}_{k,*} + (1-\mu_k)\widehat{V}_{k,1}];\\ &-c^0_k(1-\theta_k) - d_k\theta_k + \beta[\theta_k\widehat{V}_{k,*} + (1-\theta_k)\widehat{V}_{k,1}]\}.\end{aligned} \tag{2.20}$$

In all the Bellman equations above the first term in braces corresponds to active investing and the second term corresponds to passive investing (i.e. not active investing).

1. **Step**. In the first step the active set is $\mathcal{S}_0 = \{\emptyset\}$, i.e. the investor does not invest actively in any of states. Firstly we want to find index value candidates.

   Lema 2.3.3. *Under active set $\mathcal{S}_0 = \{\emptyset\}$, index value candidates are*

$$\nu^{\mathcal{S}_0}_{k,*} := -c^1_k + \beta\left(\frac{-c^0_k(1-\theta_k) - d_k\theta_k}{1-\beta+\beta\theta_k}\right),$$

$$\nu^{\mathcal{S}_0}_{k,1} := -c^1_k(1-\mu_k) + c^0_k(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k + \beta(\theta_k-\mu_k)\left(\frac{-c^0_k(1-\theta_k) - d_k\theta_k}{1-\beta+\beta\theta_k}\right).$$

   Proof. For the active set $\mathcal{S}_0 = \{\emptyset\}$, when we do not actively invest in any of states, we can straightforward from the Bellman equation (2.19) obtain that $V^{\mathcal{S}_0}_{k,*} = 0$, and for not active investment that must be valid is

$$\nu^{\mathcal{S}_0}_{k,*} = -c^1_k + \beta V^{\mathcal{S}_0}_{k,1}. \tag{2.21}$$

   We need to calculate $V^{\mathcal{S}_0}_{k,1}$. Straightforward from the equation (2.20), we obtain

$$V^{\mathcal{S}_0}_{k,1} = \frac{-c^0_k(1-\theta_k) - d_k\theta_k}{(1-\beta+\beta\theta_k)},$$

after substitution for $V_{k,1}^{\mathcal{S}_0}$ we get that the equation for not active investment is satisfied $\nu \geq \nu_{k,*}^{\mathcal{S}_0}$, where $\nu_{k,*}^{\mathcal{S}_0} = -c_k^1 + \beta\left(\frac{-c_k^0(1-\theta_k)-d_k\theta_k}{1-\beta+\beta\theta_k}\right)$ is the index value candidate for state $*$.

We continue with state 1. From the equation (2.20) we get

$$\nu_{k,1}^{\mathcal{S}_0} = -c_k^1(1-\mu_k) + \mathcal{R}_k\mu_k - \beta\mu_k V_{k,1}^{\mathcal{S}} + c_k^0(1-\theta_k) + d_k\theta_k + \beta\theta V_{k,1}^{\mathcal{S}_0},$$

after substitution for $V_{k,1}^{\mathcal{S}}$ we can derive

$$\nu_{k,1}^{\mathcal{S}_0} - c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k + \beta(\theta-\mu_k)\left(\frac{-c_k^0(1-\theta_k)-d_k\theta_k}{1-\beta+\beta\theta_k}\right).$$

□

Now we want to compare $\nu_{k,*}^{\mathcal{S}_0}$ with $\nu_{k,1}^{\mathcal{S}_0}$ to find which one of them is an index value.

Lema 2.3.4. *Under policy $\mathcal{S}_0 = \{\emptyset\}$ holds that $\nu_{k,1}^{\mathcal{S}_0} \geq \nu_{k,*}^{\mathcal{S}_0}$.*

Proof. If we subtract $\nu_{k,*}^{\mathcal{S}_0}$ from $\nu_{k,1}^{\mathcal{S}_0}$ we obtain

$$c_k^1\mu_k + c_k^0(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k + \beta(\theta_k-\mu_k-1)\left(\frac{-c_k^0(1-\theta_k)-d_k\theta_k}{(1-\beta+\beta\theta_k)}\right)$$

This is always non-negative because all terms without β are non negative. For the term $\beta(\theta_k-\mu_k-1)\left(\frac{-c_k^0(1-\theta_k)-d_k\theta_k}{(1-\beta+\beta\theta_k)}\right)$ holds that $0<\beta<1$; $(\theta_k-\mu_k-1)$ is negative and $\frac{-c_k^0(1-\theta_k)-d_k\theta_k}{(1-\beta+\beta\theta_k)}$ is also negative, so together it is non-negative. Therefore, we proved the lemma (2.3.4). □

We have shown that $\nu_{k,1}^{\mathcal{S}}$ is not only the index value candidate, but it is an index itself. In the next step we include state 1 into the active set $\mathcal{S}_1$. Note that in this step we proved the statements *i)* and *ii)* from the proposition (2.3.1), what can be observed by substitution $\nu_{k,1}$ into the equation (2.20).

2. **Step**. The active set is $\mathcal{S}_1 = \{1\}$, i.e. the investor invests actively in state 1 and does not invest actively in state $*$. We have not got index value only for state $*$, therefore in this step we have only one index value candidate, which directly is an index value for state $*$

Lema 2.3.5. *Under policy $\mathcal{S}_1 = \{1\}$, index value for state $*$ is*

$$\nu_{k,*} = \frac{-c_k^1(1-\beta+\beta\mu_k)}{(1+\beta\mu_k)} + \beta\left(\frac{-c_k^1(1-\mu_k)+\mathcal{R}_k\mu_k}{1+\beta\mu_k}\right)$$

Proof. From the first step we use the equation for not active investment that must be satisfied

$$\nu_{k,*} = -c_k^1 + \beta V_{k,1}^{\mathcal{S}_1}. \tag{2.22}$$

We need to derive $V_{k,1}^{\mathcal{S}_1}$ for $\mathcal{S}_1 = \{1\}$.

Based on the Bellman equation (2.20) and under assumption that the investor invests in state 1 we derive the balance equation

$$V_{k,1}^{\mathcal{S}_1} = \beta V_{k,1}^{\mathcal{S}_1} - c_k^1(1-\mu_k) + \mathcal{R}_k\mu_k - \nu - \beta\mu_k V_{k,1}^{\mathcal{S}_1},$$

straightforwardly we obtain

$$V_{k,1}^{\mathcal{S}_1} = \frac{-c_k^1(1-\mu_k)\mathcal{R}_k\mu_k - \nu}{1-\beta+\beta\mu_k}.$$

After substitution for $V_{k,1}^{\mathcal{S}_1}$ into (2.22) it leads us to

$$\nu_{k,*}^{\mathcal{S}_1} = \frac{-c_k^1(1-\beta+\beta\mu_k)}{(1+\beta\mu_k)} + \beta\left(\frac{-c_k^1(1-\mu_k)+\mathcal{R}\mu_k}{1+\beta\mu_k}\right)$$

for which the condition $\nu \geq \nu_{k,*}^{\mathcal{S}_1}$ is valid. Because in this step we have only one index value candidate, it is the index value $\nu_{k,*}^{\mathcal{S}_1}$. □

Similarly as in the previous step we now proved the statements *iii)* and *iv)* from the proposition (2.3.1), what is the consequence of the obtained index and equation (2.19).

2.3.5 Index rule for the original problem model 3

The original problem (P3) requires to invest money to exactly one company at a time instant t. We propose to invest money in a time instant t, to the industry $\widehat{k}(t)$ with the highest actual index, i.e.,

$$\widehat{k}(t) := \arg\max_{k\in\mathcal{K}} \nu_{k,X_k(t)}.$$

Let us recall the indices. For $0 < \beta < 1$, the index value for industry k is one of the following three index values depending on state of the industry k

$$\nu_{k,*} = \frac{-c_k^1(1-\beta+\beta\mu_k)}{(1+\beta\mu_k)} + \beta\left(\frac{-c_k^1(1-\mu_k)+\mathcal{R}_k\mu_k}{1+\beta\mu_k}\right),$$

$$\nu_{k,1} = -c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k + \beta(\theta_k-\mu_k)\left(\frac{-c_k^0(1-\theta_k)-d_k\theta_k}{1-\beta+\beta\theta_k}\right).$$

Under $\beta = 1$, we obtain the time-average version of the index values for the model 3

$$\bar{\nu}_{k,*} = \frac{-c_k^1+\mathcal{R}_k\mu_k}{1+\mu_k},$$

$$\bar{\nu}_{k,1} = -c_k^1(1-\mu_k) + \mathcal{R}_k\mu_k + d_k\mu_k + \frac{c_k^0(1-\theta_k)\mu_k}{\theta_k}.$$

Finally, we just remark that $\beta = 0$ gives the myopic version of the index values for the model 3

$$\tilde{\nu}_{k,*} = -c_k^1,$$

$$\tilde{\nu}_{k,1} = -c_k^1(1-\mu_k) + c_k^0(1-\theta_k) + \mathcal{R}_k\mu_k + d_k\theta_k.$$

Note that myopic version of the index values is the same as in the model 1.

2.4 Summary of all VCs investments models

In this chapter we proposed three different models in the restless bandits framework describing VCs investments into entrepreneurial companies. A brief overview of these models could be found in table (2.1).

Table 2.1: An overview of VCs investments models

	Model 1	Model 2	Model 3
Industry representation	More companies in industry	More companies in industry	Single company in industry
Active action (a=1)	to invest actively	to invest = to include in the portfolio	to invest actively
Passive action (a=0)	not to invest actively (passive investment)	not to invest = to exclude from the portfolio	not to invest actively (passive investment)
Set of states ($\mathcal{N}_k$)	$\{*, 0, 1\}$	$\{0, 1\}$	$\{*, 1\}$
Number of investor's investments per time period	1	M on average	1

It is important to mention that this was not done before, what is also the reason why we proposed three models and not only one. A particular investor is likely to focus on a specific goal so he can choose which of the proposed models is the most suitable. The restless bandits model allowed us to dispose of Sorensen's restrictive assumption that investing into the one industry is informative only about this industry. The reason is that indices for each company/industry change according to our action and compete against each other. If we invest in one particular company the index evolution is affected by it and in the following step it is influenced by the chosen company/industry. In all of the models we can invest on the industry level, but it is natural mainly for model 3. Model 1 and model 2 can be used for investing to particular companies. In the third model we replace the assumption that there are no arrivals, by allowing companies to be available for another investment after they have succeeded or bankrupted.

The indices for models are different from each other. At first it is hard to find similarities. But we can observe that the main idea is to evaluate the return from the investment and combine it with the future evolution by different incorporation of discount factors. Interesting is the second model, because in the situations which could be described by it, it is not important to take into account the future evolution and discount factor. The indices look to be rather complicated. It is caused mainly by considering different costs for active investments c_k^1 and inactive investments c_k^0; the same is true for bankruptcy penalties d_k^1 and d_k^0. Thus we finish this chapter by presentation of the indices in which $d_k^1 = d_k^0 = 0$ and $c_k^1 = c_k^0 = c_k$.

Model 1

$$\nu_{k,0}^{M1} = 0$$

$$\nu_{k,*}^{M1} = \frac{-c_k + \beta \mathcal{R}_k \mu_k}{(1 + \beta \mu_k)}$$

$$\nu_{k,1}^{M1} = c_k(\mu_k - \theta_k) + \mathcal{R}_k \mu_k (1 - \beta + \beta \theta_k)$$

Model 2

$$\nu_{k,0}^{M2} := 0$$

$$\nu_{k,1}^{M2} := -c_k(1 - \mu_k - \theta_k) + \mathcal{R}_k \mu_k$$

Model 3

$$\nu_{k,*}^{M3} = \frac{-c_k^1 + \beta \mathcal{R}_k \mu_k}{(1 + \beta \mu_k)} = \nu_{k,*}^{M1}$$

$$\nu_{k,1}^{M3} = \mathcal{R}_k \mu_k + \frac{c_k(\mu_k - \theta_k)}{1 - \beta + \beta \theta_k}$$

Chapter 3

Partially observable Markov decision processes and learning methods

The goal of the thesis is to deal with uncertainty in the VCs investments by adopting learning into the proposed solution. The first two chapters introduced the multi-armed restless bandits framework and three different models for describing VCs investments in entrepreneurial companies are introduced there as well. The aim of this chapter is to describe the most important approaches and frameworks necessary for the incorporation of learning. To this end we refer to partially observable Markov decision processes. Their definition is followed by a description of the Bayesian updating and the chapter is finished by a short overview of a behavioural study which validates the capability of the Bayesian updating in combination with the multi-armed restless bandits to describe reality adequately.

3.1 Partially observable Markov decision processes

Partially observable Markov decision processes (POMDPs) generalize MDPs (introduced in section (1.1)). The underlying system is an MDP, but we cannot observe the exact value of the state in a time instant t, which we denote n_t. On the other hand, we observe the noise-corrupted partial information about the system state in each time instant, which we denote x_t. Further by a we denote the chosen action and by $0:t$ observations received up to instant t. We can reformulate this problem as a fully observable MDP in which decision state is the conditional probability distribution of system state conditioned on previously received observations (see Williams (2007)).

Let us denote,

- $\mathbb{N}_t = p(n_t|x_{0:t-1}; a_{0:t-1})$ - the conditional probability distribution of the system state,
- $g_t(n_t, a_t)$ - the reward per period,
- $g_T(n_T)$ - the terminal reward.

Given the conditional probability distribution of the system state we can calculate the expected value of the reward

$$g_t(\mathbb{N}_t, a_t) = \sum_{n_t} g_t(n_t, a_t) p(n_t | x_{0:t-1}; a_{0:t-1}),$$

$$g_T(\mathbb{N}_T) = \sum_{n_T} g_T(n_T) p(n_T | x_{0:T-1}; a_{0:T-1}),$$

so the reward per state and terminal reward can be expressed as functions of the underlying system, what is one of the fundamental POMDPs assumptions. POMDP solution policy prescribes the optimal action for each possible belief state for all possible states. Optimal policy is the sequence of optimal actions and optimal policy maximizes the expected reward over an infinite horizon.

POMDPs are P-SPACE hard (i.e. as hard as any problem which is solvable using an amount memory that is polynomial in the problem size, and unlimited computation time (Blondel and Tsitsiklis (2000) and Papadimitriou and Tsitsiklis (1987))). The study by Michael L. Littman and Kaelbling (1995) found the solution times in order of hours for the problem with fifteen underlying system states and observations values and four actions.

The POMDP modelling framework is very powerful to model a variety of real-world situations. Nowdays it is hugely used in artificial inteligence and automated planning applications. Examples of such applications are robot navigation problems, sensor management and all planning applications under uncertainty.

3.2 Bayesian updating

In this thesis we would like to optimize our investment selections under uncertainty. This is often the case in finances. POMDP is a suitable framework for description of such situations because the state process is not observed directly. Thus we have only a certain *belief* that the unobservable system is in state x_k. POMDP is Markovian, therefore after we choose the action we earn a reward and based on them we update our belief. This belief could be updated by Bayesian updating, which we introduce in this section following Pastor and Veronesi (2009).

In general, suppose that we are uncertain about an arbitrary parameter μ. At the beginning we have normally distributed prior beliefs about μ with mean μ_0 and variance σ_0^2. After observing T independent signals s_t about μ, we obtain revised posterior belief according to the Bayes' rule. The independent signals are $s_t = \mu + \epsilon_t$, where ϵ_t is normally distributed with zero mean and known variance σ^2; by $\bar{s}$ we denote the average signal value $\bar{s} = \frac{1}{T}\sum_{t=1}^{T} s_t$. Our posterior belief is then normally distributed with mean

$$\bar{\mu}_T = \mu_0 + \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{T}{\sigma^2}} + \bar{s}\frac{\frac{T}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{T}{\sigma^2}},$$

and variance

$$\bar{\sigma}_T^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{T}{\sigma^2}}.$$

This definition fulfils our expectations that learning reduces uncertainty. The uncertainty about parameter μ is the posterior variance $\bar{\sigma}_T^2$, which decreases with increasing number of observations T. An observed signal s_t that is higher than the expectation, $s_t > \bar{\mu}_{t-1}$, is rising our expectations and a signal that is smaller than its expectation is lowering our expectations.

3.3 Experimental evidence from simulating real world financial systems

At the end of this chapter we briefly present a study by Payzan-LeNestour (2012) in which the authors test experimentally the suitability of the representation of real life financial situations by restless bandits with incorporated Bayesian updating. More preciously, first they try to examine if investing in financial assets is one of the areas were humans can achieve Bayesian reasoning. Secondly, they examine it by incorporation of the multi-armed restless bandits suitable mainly for description of more structured financial instruments that trade exclusively in the over-the-counter market.

In the paper they present a six-armed restless bandits board game, where players, after having chosen an action (investing in investment opportunity) receive reward or in some situations penalty. The player uses the Bayesian updating for understanding where she should invest. They carry out their experimental study on sixty-one undergraduate students from École Polytechnique Fédérale in Lausanne.

They empirically prove that humans act in the Bayesian way, so people can be good learners in sophisticated problems. The result is supported by results from other papers (Bruguier and Bossaerts (2010)) that humans can extract informations from financial data better than computers. The paper also shows that economical performance of agents learning in the Bayesian way is much better than adaptive learners. An important conclusion for this paper is that, unlike in general, in financial applications as VCs investing they learn in a sophisticated way. This is described adequately by incorporation of the Bayesian updating into the multi-armed restless bandits problems.

Chapter 4

Learning venture capitalists investments model and simulation study

In this chapter we introduce a learning model for VCs investments in entrepreneurial companies based on model 3 from the first chapter. We have chosen model number 3, because we believe that it has the best descriptive power for the real VCs investments. Recall that in model 3 we solve the problem with the option to invest in the same industry more than once, on the other hand we assume that we invest in the industry level. This restriction has the advantage that it helps us to propose a model which could be confronted with the results in Sorensen (2008).

We incorporate uncertainty about the success probability, because we assume that it is not observable. Therefore we define the model in the partially observable MDPs environment (POMDPs) and we use the Bayesian updating for revising our belief about success probability.

In this chapter we first describe the model and define it as an POMDP. Next we introduce the theoretical background for the simulation study. We show other reference strategies and at the end we present results of an extensive simulation study.

4.1 Learning VCs investments model

4.1.1 Problem description learning VCs investments model

At each time period the investor (VC) chooses among K possibilities where to invest, denoted $k = 1, 2, ..., K-1$. Each possibility is an industry that is represented by an entrepreneurial company belonging to this industry. For instance such industries could be: Health/Biotechnology, Communications/Media, Computer Hardware/Electronics, Software, Consumer/Retail etc. The possibility K represents the option of not investing in any company.

The outcome of an investment is *success* (W) or *failure* (F), as indicated by $y_k(t) \in \{W, F\}$. The success probability μ_k of each company is unknown and unobservable. We assume that this probability is constant over time, but can be different for various industries. In the case of the known company's state this probability is given as $\mu_k = Pr[y_k(t) = W]$. Not knowing μ_k, the investor has a prior belief with distribution $\mathbb{D}_k(0)$. The distribution $\mathbb{D}_k(t)$ varies with time. We represent investor's belief as the empirical mean of the distribution $\mathbb{D}_k(t)$ and we denote it $x_k(t)$. Therefore the prior belief is given by $x_k(0)$. After each investment, the investor is updating this belief using the Bayes rule. The support of $x_k(t)$ is the interval $[0, 1]$, representing all possible values of μ_k.

The binary outcomes help to simplify the Bayesian updating process. We can assume that the investor's initial beliefs are β-distributed, $\mathbb{D}_k(0) = Beta(u_k(0), v_k(0))$. Let $j_k(t)$ be the number of past successes and $l_k(t)$ be the total number of past investments in industry k and a time instant t. The updated beliefs are then simply $Beta(u_k(t), v_k(t))$,where $u_k(t) = u_k(0) + j_k(t)$ and $v_k(t) = v_k(0) + l_k(t) - j_k(t)$. Therefore we can say that $u_k(t)$ corresponds to the number of past successes and $v_k(t)$ corresponds to the number of past failures. As the number of investments increases, the mass of the distribution becomes concentrated at the empirical success rate, defined as $x_k(t) = \frac{u_k(t)}{u_k(t)+v_k(t)}$, which also equals the mean of the $Beta(u_k(t), v_k(t))$ distribution. In other words, given the investor's beliefs, $x_k(t)$ is the expected value of μ_k.

In the investing process we look only on successes W and failures F. In the following we incorporate also the possibility that the company *hibernates* and *resurrects*. We denote it H and R respectively. H can happen when the investor is active but the company will not succeed in that particular time instant. In such a case our belief changes and it falls. The reason is that the time delay in becoming successful cause the investor additional costs, thus it lowers our perspective of the value of the firm. R can take place when the investor is passive and the company does not get bankrupt in that particular time instant. In this case our belief rises, because we experienced that the company is viable. After the company bankrupts the success probability obviously falls. The success probability rises only when we are active concerning the company and the time instant ends with company's success.

To summarize, if the investor invests in the company k, when being in the belief state $x_k(t)$, then at the end of the period the investor receives feedback:

$$o_{k,x}(t) = \begin{cases} W & \text{w.p. } x_k(t) \\ H & \text{w.p. } (1 - x_k(t)) \end{cases}$$

If the investor does not invest in the company k when it is in the belief state $x_k(t)$, then at the end of the same period the investor receives feedback:

$$o_{k,x}(t) = \begin{cases} F & \text{w.p. } \theta_k \\ R & \text{w.p. } (1-\theta_k) \end{cases}$$

In the updating process the number of past successes $u_k(t)$ and failures $v_k(t)$ varies. Therefore, we update the distribution of our belief $\mathbb{D}_k(t)$. Since $u_k(t)$ and $v_k(t)$ grow with t and, therefore, take an unbounded number values, we have to bound the maximum number of events. Let M be the maximum number of past successes and failures. Then the interval $[0, 1]$ of all possible success probability values is partitioned into a finite number of intervals. When at least one of the $u_k(t)$ or $v_k(t)$ is equal to $M+1$ we have to round it. We round our belief to the closest lower belief in which the parameter that was $M+1$ is now equal to M and the other parameter is lower or equal to M.

Another feature we should be aware of is that after the company's success (initial public offering (IPO) or acquisition) or after its bankruptcy we set our new prior belief to be equal to the last belief for this particular industry.
We do not incorporate arrivals into this model, therefore it falls into the multi-armed restless bandit framework. By letting the company transition from a success or a failure to the original state, the model recovers arrivals. Because success probability is not observable we make a POMDP formulation of this problem.

4.1.2 POMDP formulation of learning VCs investments model

We denote by $\mathcal{A} := \{0, 1\}$ the *action space*. Here, action $a := 0$ means that the investor does not invest actively in the company, and action $a := 1$ means that the investor invests actively in the company. This action space is the same for every company k.

Each company/industry k is defined independently of the other companies/industries as the tuple

$$\left(\mathcal{N}_k, (\boldsymbol{W}^a_k)_{a\in\mathcal{A}}, (\boldsymbol{R}^a_k)_{a\in\mathcal{A}}, (\boldsymbol{P}^a_k)_{a\in\mathcal{A}}\right),$$

where

- $\mathcal{N}_k := * \cup \mathcal{X}_k$ is the *state space*.

  State $* := [n^*_1, n^*_2]$ represents the prior belief distribution parameters about the company without any investments or about the company representing an industry which already belongs to the investor's portfolio but it has already succeeded or bankrupted; and $\mathcal{X}_k := [n_1, n_2]$ is the belief state distribution (posterior) of the k^{th} company's success probability. In other words, if the company is in the state $\mathcal{X}_k := [n_1, n_2]$ it means that the company is in the current investor's portfolio. It holds that $n^*_1, n^*_2, n_1, n_2 \leq M$, where M is the biggest possible number of past successes and failures.

When the company is in the state $\mathcal{X}_k$ than our belief is represented by $x_k = \frac{n_1}{n_1+n_2}$, similarly when the company is in the state $*$ the belief is represented by $x_k = \frac{n_1^*}{n_1^*+n_2^*}$. When the company is going from the state $\mathcal{X}_k$ to state $*$ we set $[n_1^*, n_2^*] := [n_1, n_2]$.

- $\boldsymbol{W}_k^a := \left(W_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $W_{k,n}^a$ is the (expected) one-period attention consumption required by company k at state n if action a is decided at the beginning of a period; in particular, for any $n = [n_1, n_2] \in \mathcal{N}_k$,

$$W_{k,n}^1 := 1, \qquad\qquad W_{k,n}^0 := 0;$$

- $\boldsymbol{R}_k^a := \left(R_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $R_{k,n}^a$ is the expected one-period *reward* earned by the investor for company k at state $n = [n_1, n_2] \in \mathcal{N}_k$ if action a is decided at the beginning of a period; in particular,

$$\begin{aligned} R_{k,[n_1^*,n_2^*]}^1 &:= -c_k^1,\ R_{k,[n_1^*,n_2^*]}^0 := 0, \\ R_{k,[n_1,n_2]}^1 &:= -c_k^1 \cdot (1 - x_k(t)) + \mathcal{R}_k x_k(t), \\ R_{k,[n_1,n_2]}^0 &:= -c_k^0(1-\theta_k) - d_k\theta_k \end{aligned}$$

Where $\mathcal{R}_k > c_k^1 > c_k^0$,$d > c_k^0$, and $x_k(t) = \frac{n_1}{n_1+n_2}$ or $x_k(t) = \frac{n_1^*}{n_1^*+n_2^*}$

- $\boldsymbol{P}_k^a := \left(p_{k,n,m}^a\right)_{n,m\in\mathcal{N}_k}$ is the company k stationary one-period *state-transition probability matrix* if action a is decided at the beginning of a period, i.e., $p_{k,n,m}^a$ is the probability of moving to state $m = [m_1, m_2]$ from state $n = [n_1, n_2]$ under action a; in particular, we have

$$p_{k,[n_1^*,n_2^*],[n_1^*,n_2^*]}^0 = 1$$

$$p_{k,[n_1^*,n_2^*],[n_1,n_2]}^1 = 1$$

$$p_{k,[n_1,n_2],m}^0 = \begin{cases} \theta_k & \text{if } m = [m_1^*, m_2^*] \text{ ,where } [m_1^*, m_2^*] := [n_1, n_2 + 1] \\ 1 - \theta_k & \text{if } m = [n_1 + 1, n_2] \end{cases}$$

$$p_{k,[n_1,n_2],m}^1 = \begin{cases} x_k(t) & \text{if } m = [m_1^*, m_2^*] \text{ ,where } [m_1^*, m_2^*] := [n_1 + 1, n_2] \\ (1 - x_k(t)) & \text{if } m = [n_1, n_2 + 1] \end{cases}$$

The dynamics of the company k is thus captured by the state process $n_k(\cdot)$ and the action process $a_k(\cdot)$, which correspond to state $n_k(t) \in \mathcal{N}_k$ and action $a_k(t) \in \mathcal{A}$

at all time instants $t \in \mathcal{T}$. As a result of deciding action $a_k(t)$ in state $n_k(t)$ at a time instant t, the company k uses the allocated attention, earns the reward, and evolves its state for the time instant $t+1$. If $n_k(t) \in \mathcal{N}_k$, then the state evolution is the same as the belief's distribution evolution.

To summarize:

- If the investor does not invest in the company k at time t ($a_k = 0$), the evolution of the investor's belief has the distribution:

$$\mathbb{D}_k(t+1) = \begin{cases} Beta(n_1(t), n_2(t)+1) \text{ for } o_{k,x}(t) = F \\ Beta(n_1(t)+1, n_2(t)) \text{ for } o_{k,x}(t) = R \end{cases}$$

- If the investor invests in the company k at time t ($a_k = 1$), the evolution of the investor's belief has the distribution:

$$\mathbb{D}_k(t+1) = \begin{cases} Beta(n_1(t)+1, n_2(t)) \text{ for } o_{k,x}(t) = W \\ Beta(n_1(t), n_2(t)+1) \text{ for } o_{k,x}(t) = H \end{cases}$$

which reflects the Bayesian updating in the company/industry k after observing success, failure, hibernation or resurrection.

4.1.3 Optimization problem: learning VCs investments model

After the model definition, we define the optimization problem following the notation introduced in the chapter 2.

The problem is to find a joint policy π maximizing the objective given by the discounted aggregate reward starting from the initial time instant 0 subject to the family of sample path allocation constraints, i.e.,

$$\begin{aligned} &\max_{\boldsymbol{\pi} \in \Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}_0^{\boldsymbol{\pi}} \left[\sum_{k \in \mathcal{K}} R_{k,X_k(\cdot)}^{a_k(\cdot)} \right] \\ &\text{subject to } \left[\sum_{k \in \mathcal{K}} a_k(t) \right] = 1, \text{ for all } t \in \mathcal{T}. \end{aligned} \tag{BP}$$

4.1.4 Relaxation and decomposition

Problem (BP) can be relaxed by requiring to invest in one company only on average as proposed in Whittle (1988), which is further approached by incorporating a Lagrangian multiplier ν and it can be decomposed into a parameterized optimization

problem below. Notice that any joint policy $\pi \in \Pi_{\mathbf{X},a}$ defines a set of single-company policies $\bar{\pi}_k$ for all $k \in K$, where $\bar{\pi}_k$ is a randomized and non-anticipative policy depending on the joint state-process $\mathbf{X}(\cdot)$ and deciding the company-k action process $\mathbf{a}(\cdot)$. We write $\bar{\pi}_k \in \Pi_{\mathbf{X},a_k}$ and we therefore study the company-k subproblem

$$\max_{\widetilde{\pi}_k \in \Pi_{\boldsymbol{X},a_k}} \mathbb{B}_0^{\widetilde{\pi}_k} \left[R_{k,X_k(\cdot)}^{a_k(\cdot)} - \nu W_{k,X_k(\cdot)}^{a_k(\cdot)} \right]. \tag{4.1}$$

The main idea of our approach is to identify a set of optimal policies $\bar{\pi}_k^*$ for (4.1) for each $k \in K$, and using them to construct a joint heuristic policy π, feasible though not necessarily optimal for problem (BP).

4.1.5 Optimal solution to single company subproblem

As we explained in the chapter 2 and in the appendix (A. 1), in certain cases, problem (4.1) can be optimally solved by assigning a set of index values $\nu_{k,n}$ to each state $n \in \mathcal{N}_k$ (Niño-Mora (2007); Jacko (2010a)). If this is the case, the problem is called indexable. Therefore we propose following conjecture

Conjecture 4.1.1. *(Indexability) The problem* (4.1) *is indexable.*

In contrast with the models presented in the chapter 2, this learning model falls into the real-state multi-armed restless bandits. Therefore, computation in closed-form of the Whittle index is ruled out for this model. We exploit the algorithm used in Villar (2012) for calculating Whittle index. In Villar (2012) they focus mainly on the applications on elusive targeting and multi-target tracking. To make the algorithm valid for financial applications we have done several modifications. Below we briefly introduce the main idea behind the algorithm.

Algorithm for calculating Whittle index for the real-state multi-armed restless bandits

We use notation that is described more extensively in the appendix (A. 2) and in (A. 3). For a given active set $\mathcal{S}$, we denote by $\langle a, \mathcal{S} \rangle$ the policy that choose action $a \in \{0, 1\}$ in a initial time instant and adopts the $\mathcal{S}$-active policy thereafter. For a time instant t using the notation $n = [n_1(t), n_2(t)]$ and $\vec{n} = [n_1(t+1), n_2(t+1)]$, we define *marginal reward measure*

$$R_n^{\langle 1,S \rangle} - R_n^{\langle 0,S \rangle} = R_n^1 - R_n^0 + \beta \sum_{\vec{n}} [p_{n,\vec{n}}^1 - p_{n,\vec{n}}^0] R_{\vec{n}}^S,$$

and *marginal work measure*

$$W_n^{\langle 1,S \rangle} - W_n^{\langle 0,S \rangle} = W_n^1 - W_n^0 + \beta \sum_{\vec{n}} [p_{n,\vec{n}}^1 - p_{n,\vec{n}}^0] W_{\vec{n}}^S.$$

Then based on Niño-Mora (2002, 2006) for the company k in state $n_k \in \mathcal{N}_k$ we can calculate the Whittle index by means of the formula (see appendix (A. 2) especially equation (A.6))

$$\nu_{k,n} = \frac{R_{n_k}^{\langle 1,S\rangle} - R_{n_k}^{\langle 0,S\rangle}}{W_{n_k}^{\langle 1,S\rangle} - W_{n_k}^{\langle 0,S\rangle}}, \tag{4.2}$$

In the simulation algorithm, we use the formula above to calculate the index for each company k and for every state $n_k \in \mathcal{N}_k$. Because the equations for marginal reward measure and marginal work measure is the certain sum through all the possible future states we have to simulate each index value more than 10^4 times and as the index value we select the mean from obtained values.

4.1.6 Heuristic rule for the original problem

The original problem (BP) prescribes to invest exactly in one company at a time instant t, therefore we need to propose a heuristic rule for it.

Rule 1. *At any time instant t the investor has to invest in the company $\widehat{k}(t)$ with the highest actual index*

$$\widehat{k}(t) := \arg\max_{k\in\mathcal{K}} \nu_{k,X_k(t)}.$$

4.2 Simulation study

We study the performance of our proposed heuristic rule by a number of different simulations. In the simulations we first calculate the indices as proposed above. Then we simulate possible evolutions under our heuristic rule and under alternative rules. We have carried out simulations for more than one hundred different scenarios and for each scenario we simulate it 10^3 times. In the figures we show the average from the obtained values with confidence intervals. In this part of the thesis, we present only a representative sample of all simulations. The simulation study is divided into three sections: alternative rules, β-study and selection from various scenarios.

4.2.1 Alternative rules

Because investment strategies of real VCs are not publicly known, we compare our heuristic rule mainly against usually assumed strategies, where investor chooses to invest in the company with the highest return on investment (ROI). Unfortunately, there is not any official definition of ROI in the multi-armed restless bandits environment. Usually ROI is calculated as a ratio of the net return and the cost of an investment. Thus we propose *basic ROI* rule for the company k in the multi-armed restless environment as

$$ROI = \frac{(\mathcal{R} - c_k^1) - (c_k^0 - d_k)}{c_k^1 - c_k^0}. \tag{4.3}$$

The second rule is the ROI in which we try to incorporate also probabilities. The success probability is unknown and for observability of the improvement caused by our learning we do not want to incorporate our belief $x_k(t)$ into the formula, thus we incorporate only θ_k into the *stochastic version of ROI*

$$SROI = \frac{(\mathcal{R} - c_k^1) - (c_k^0(1-\theta_k) - d_k\theta_k)}{c_k^1 - c_k^0\theta_k}. \tag{4.4}$$

For evaluation of the learning significance we define the *unobservable version of ROI* (the UNROI), in which we use the probabilities that are initially unknown for our index solution. This comparison could show us what is the "price" of the uncertainty.

$$UNROI = \frac{(\mathcal{R}\mu_k - c_k^1(1-\mu_k)) - (c_k^0(1-\theta_k) - d_k(\theta_k))}{c_k^1(1-\mu_k) - c_k^0(1-\theta_k)}. \tag{4.5}$$

The last alternative rule is the closed-form Whittle index for model 3 in the chapter 2. This rule also uses the unobservable parameters. The model 3 index value

- for company k in state $*$ is

$$\nu_{k,*} = \frac{-c_k^1(1-\beta+\beta\mu_k)}{(1+\beta\mu_k)} + \beta\left(\frac{-c_k^1(1-\mu_k) + \mathcal{R}_k\mu_k}{1+\beta\mu_k}\right) = \frac{-c_k^1 + \beta\mathcal{R}_k\mu_k}{1+\beta\mu_k},$$

- for company k in state 1 is

$$\nu_{k,1} := -c_k^1(1-\mu_k) + \mathcal{R}\mu_k + \left[c_k^0(1-\theta_k) + d_k\theta_k\right] \cdot \left(\frac{1-\beta+\beta\mu_k}{1-\beta+\beta\theta_k}\right). \tag{4.6}$$

In the simulations we consider six different companies by which we try to emulate real industries. The majority of simulations are based on the parameters setting which is described in the table (4.1).

The parameters for c^1, c^0, d and $\mathcal{R}$ are in millions of EUR, for instance. Therefore, the total reward in the figures is also in millions of EUR. We prescribe success probability and variance with which we generate the investors observations. The discount factor is assumed to be the same for all companies. In the following figures we vary success probability of a company 6 from 0.3 to 0.8. Convention is that the left figure shows total earned reward by the investor and the right figure shows the relative gap between our rule and the alternative rule.

Return of investments In figures (4.1) and (4.2) we compare the resulting Whittle index rule against ROI (4.3). The proposed rule is significantly better and the suboptimality gap is between 35% and 65% for different values of the success probability. Therefore we can claim that the whittle inder rule outperforms ROI in this setting.

Table 4.1: Parameters used in the simulation study

	c^1	c^0	d	$\mathcal{R}$	θ	μ	σ^2	β
Company 1	0.003	0.001	0.01	0.989	0.3	0.6	0.1	0.9
Company 2	0.08	0.009	0.01	0.5	0.7	0.3	0.3	0.9
Company 3	0.01	0.01	0.0001	0.35	0.45	0.45	0.2	0.9
Company 4	0.02	0.02	0.02	0.1	0.5	0.5	0.5	0.9
Company 5	0.001	0.003	0.1	0.5	0.6	0.1	0.35	0.9
Company 6	0.003	0.001	0.01	0.989	0.3	0.6	0.1	0.9

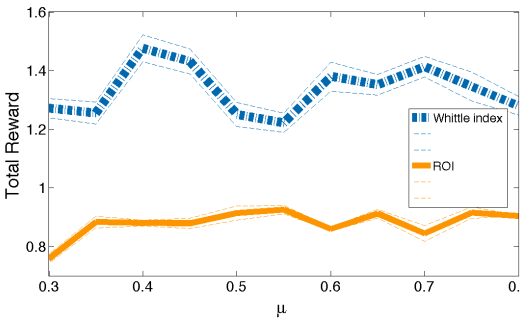

Figure 4.1: ROI - Total reward

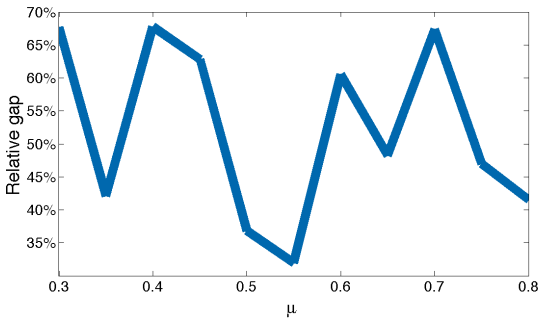

Figure 4.2: ROI - Relative gap between rules

Stochastic return of investments Results are similar as obtained for ROI. In this example SROI is worse than ROI, but that is not true for all other settings. Having gone through examples we observed that we can easily outperform ROI, but with SROI it is much harder. That leads us to use SROI for comparison in the other sections.

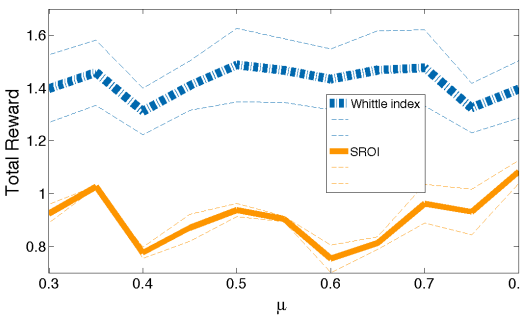

Figure 4.3: SROI - Total reward

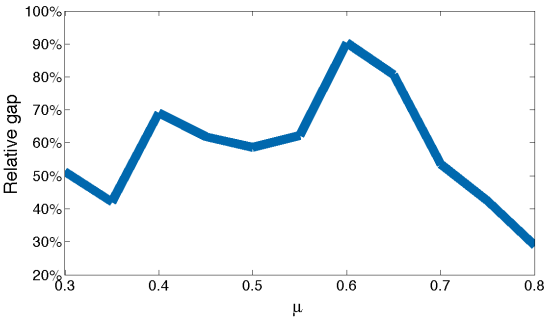

Figure 4.4: SROI - Relative gap between rules

Unobservable version of ROI Following rule is UNROI (figures (4.5) and (4.6)) which includes the unobservable success probability. That is one of the main reasons that our rule is not as good as before. We can say that our rule and UNROI performance is similar, but we must keep in mind of broad confidence intervals.

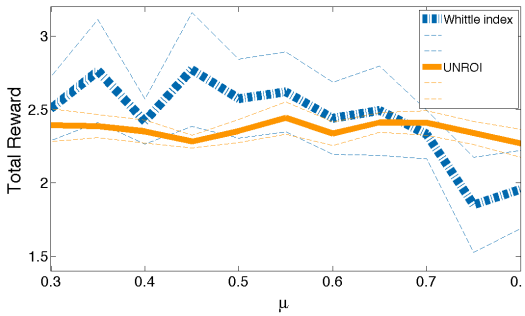

Figure 4.5: UNROI - Total reward

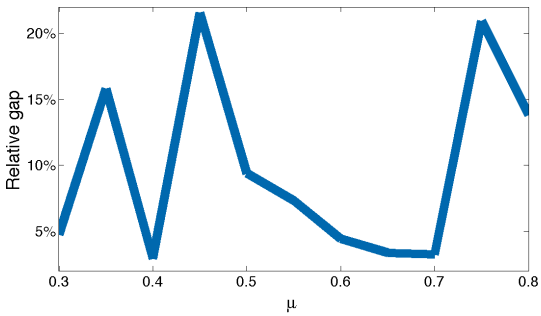

Figure 4.6: UNROI - Relative gap between rules

Model 3 index The last alternative rule considered is index for model 3. As in the previous case the alternative rule has the advantage that it knows the probability. In the figures (4.7) and (4.8) we can observe that alternative rule is much better than our proposed rule. On the other hand, situations when the success probability is high enough our rule suddenly becomes better. It seems that our rule is better in identifying the company with high probability than index from model 3.

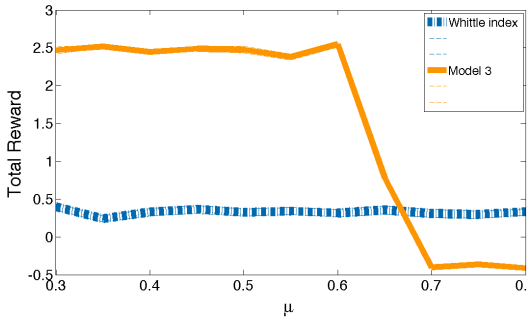

Figure 4.7: Model 3 index - Total reward

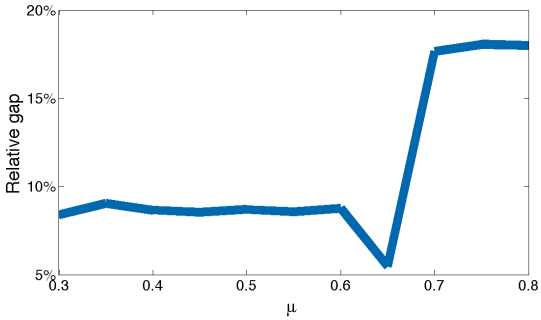

Figure 4.8: Model 3 index - Relative gap between rules

In the rest of the simulations we use SROI rule, because based on our simulations we can say that it competed better with our proposed rule than ROI. The results of comparison with other rules show us that if the rule does not know the probability we can become much better than alternative rules. On the other hand the result of the last simulation suggests that our learning can be much better as it is now. We are aware that we have not sufficient knowledge of investment strategies used by investors in practice. Neither we know the optimal strategy. Therefore we try to do as much as we are able, but we must keep in mind that in reality our performance could be worse in comparison with real strategies.

4.2.2 β-study

Discount factor is very important for financial simulations. Therefore we are interested how our rule react on different values of β. In the following four pictures we can

observe performance for four different dicount factor values: 0.6, 0.8, 0.9 and 0, 99 respectively (corresponding figures (4.9),(4.10), (4.11) and (4.12)).

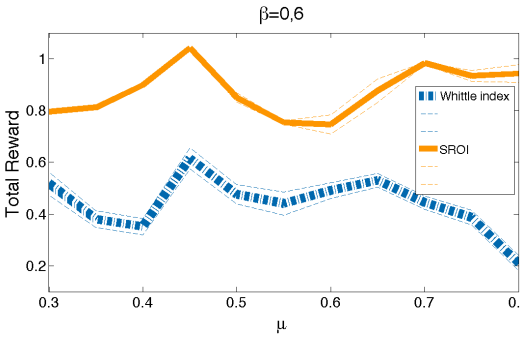

Figure 4.9: $\beta = 0,6$ - Total reward

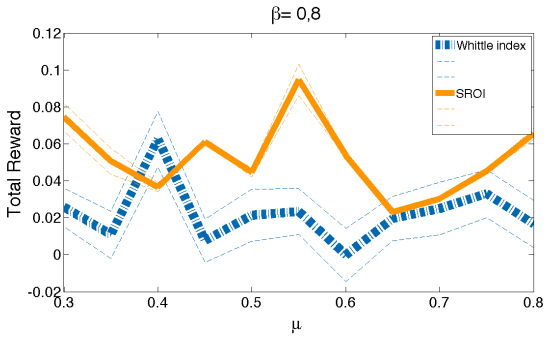

Figure 4.10: $\beta = 0,8$ - Total reward

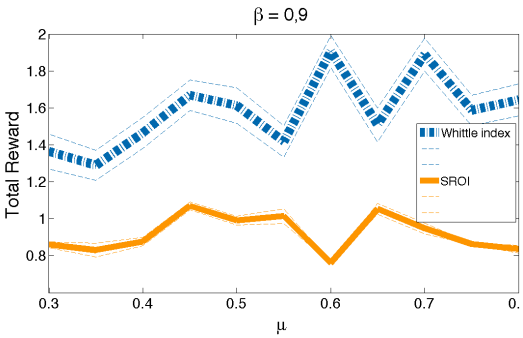

Figure 4.11: $\beta = 0,9$ - Total reward

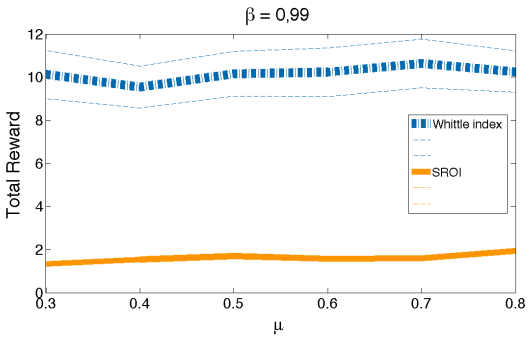

Figure 4.12: $\beta = 0,99$ - Total reward

Results of β-study are interesting. As we can see our rule is becoming better with rising β, but for small values of β SROI outperforms our rule. As we know β in finances is often between $0,9$ and 1 therefore it suggests that our rule is more suitable for financial applications than SROI.

4.2.3 Different parameters selections

In the last simulation section we show a sample of possible changes in parameters. In the first simulation we vary variance for the company 6.

In the figure for the total reward (4.13) we can observe that the limits of the confidence interval increases. The reasons for this are two. First for the fixed number of simulations our ability of learning is lowering with an increasing variance. A second point is that our proposed rule does not include the variance. Therefore, our rule is not considering rising variance. We think that this is the main weakness of the proposed rule and it should be our focus in further research.

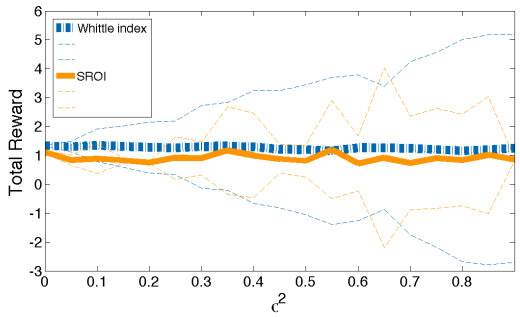

Figure 4.13: Variance simulation - Total reward

We have carried out also many other changes in parameters, but we do not present them here. There is infinite number of possible settings for companies. A majority of our results suggests that our rule outperforms SROI as in the figure (4.9) or the worse scenarios are similar to the figure (4.5) or as the figure (4.10). We believe that this evidence is promising and it should provide an importance for more studies based on the proposed methodology.

Conclusion

In this work we study mathematical methods for description of venture capitalists investments in entrepreneurial companies. Such investments are filled by uncertainty about technologies and investment opportunities. Our work is based on the results obtained in Sorensen (2008). An econometric study based on the classical multi-armed bandits and Bayesian updating by Sorensen shows that VCs investment decision is based on expected return from the investment itself and on the potential to learn from it. Moreover, he rejected the hypothesis that individual investments are made in isolation. On the other hand, Sorensen was forced to establish several not fully reasonable restrictions. One of the main improvements of our work is the replacement of the classical multi-armed bandits by the restless bandits. It allows us to avoid restrictions that investors learn only from their own past investments and that investments in one industry are not informative about investments in other industries.

Application of the multi-armed restless bandits framework for description of financial applications is challenging and also innovative, because up to this moment it was mainly used only in the communication networks. Therefore we first introduce the theoretical background and then we propose three different multi-armed restless bandits models for description of VCs investments. Sorensen (2008) did not describe dynamics of such investments, thus we try to show various ways of description. To the best of our knowledge this is the first time VCs investments are captured by multi-armed restless bandits. For each model we derive the closed-form Whittle index and we propose a heuristic rule based on the obtained indices. For index derivation we analytically follow the AG-algorithm proposed by Niño-Mora (2007). Every model avoids some assumptions from Sorensen (2008). For instance, in model 1 and 2 the investor does not need to invest in the industry level. Moreover, in the model 2 she can simultaneously invests in M companies. Model 3 allows to invest in the same industries more than once.

Based on the model 3 we developed a learning model describing VCs investments. We formulate the problem as a partially observable Markov decision process, because the success probability of the company is not observable. Therefore the investor has only a certain belief of success for the particular company. According to successes

and failures in investments he updates this belief in the Bayesian way. The combination of the multi-armed restless bandits and the Bayesian updating was shown to be promising and suitable for description of real financial situations by the behavioural study done by Payzan-LeNestour (2012) (see section (3.3)).

The learning VCs investment model falls into the multi-armed restless bandits with continuum of states, therefore instead of deriving a closed-form Whittle index, we obtain the index by simulations. We exploit the algorithm used in Villar (2012). To study the performance for this Whittle index policy we refer to an exhaustive simulation study. We study more than one hundred different scenarios. In the section (4.2) dedicated to the simulation study we show representative sample consisting from three parts. First we show the differences between different rules and also the price of uncertainty. Second part studies the changes of performance for different discount factors and the last section study the differences caused by parameter variations.

The simulation study suggests that our rule is well performing in comparison with other rules. On the other hand we identified several areas for possible improvements in particular it is difficult to compare our rule with the results of strategies used in practice. For the further work it is interesting to include into index also variance of our belief. Moreover, by reformulation of probabilities it seems that it is possible to connect our index with the table values for the Gittins index (i.e. classical multi-armed bandits index), which would allowed us to compare ours with Sorensen's results by the same econometric study.

Resumé

V práci skúmame metódy vhodné na opis investovania investorov s rizikovým kapitálom do firiem, menších podnikov a podnikov mimo burzy. Pri takýchto investíciách investor čelí technologickým neistotám a neistotám ohľadom možných investičných príležitostí. Táto práca je postavená predovšetkým na výsledkoch článku Sorensen (2008). Sorensen urobil ekonometrickú štúdiu za použitia "classical multi-armed bandits" spolu s Bayesovskou aktualizáciou a vďaka nej ukázal, že investori investujú na základe návratnosti investície a na základe možnosti učiť sa z danej investície. Inak povedané, investor môže uprednostniť investíciu do firmy, od ktorej neočakáva ziskovosť, ale ktorá mu poskytne vedomosti o nových technológiách a informácie o novom odvetví. Okrem toho Sorensen zamietol hypotézu, že investície sú robené izolovane. Na druhej strane bol Sorensen prinútený predpokladať niekoľko nie úplne vhodných obmedzení pre finančné aplikácie. Jedným z hlavných zlepšení dosiahnutých v tejto diplomovej práci je, že sme nahradili "classic bandits" pomocou "restless bandits". V "restless" verzii sa firmy vyvíjajú do nových stavov aj keď do nich práve investor neinvestuje. Preto sme nemuseli predpokladať, že investori sa učia len z vlastných minulých investícií a že investície do jedného odvetvia neposkytujú informáciu o investovaní do iných odvetví.

"Multi-armed restless bandits" boli doposiaľ používaní predovšetkým v telekomunikačných aplikáciách. Preto ich použitie vo financiách bolo inovatívne ale tým pádom sme museli vyriešiť niekoľko netriviálnych problémov, ktorými sa doposiaľ nikto nezaoberal. V práci najprv predstavujeme základné princípy a charakteristiky použitých metód. Následne vytvoríme tri rôzne modely popisujúce investovanie rizikového kapitálu. Pre každý z nich odvodíme Whittlov index, na základe ktorého vytvoríme heuristické pravidlo pre investovanie. V prvom a treťom modeli Whittlov index odvodíme pomocou AG-algoritmu, navrhnutého v Niño-Mora (2007). Keďže Sorensen vďaka ekonometrickej štúdii nebol nútený modelovať dynamiku takýchto investícií, tri dané modely ukazujú rôzne pohľady, ako takéto investovanie naformulovať. Na základe našich znalostí toto je prvá práca, v ktorej je investovanie rizikového kapitálu opísané pomocou "multi-armed restless bandits". Okrem toho každý z uvedených modelov sa snaží vyhnúť niektorým predpokladom z článku Sorensen (2008). Napríklad v modeloch jedna a dva neinvestujeme len do odvetví, ale môžeme investo-

vať priamo do firiem, t.j. z jedného odvetvia môžeme mať viacero firiem. V druhom modeli máme možnosť investovať do M firiem súčasne. Nakoniec v modeli 3 môžeme do jedného odvetvia/firmy investovať opakovane.

Na základe modelu 3 následne vyvinieme učiaci sa model popisujúci investovanie rizikového kapitálu. Úlohu sformulujeme ako čiastočne pozorovateľný markovovský rozhodovací proces, pretože pravdepodobnosť úspechu firmy je neznáma. Investor má len určité presvedčenie odhadujúce šance na úspech firmy. Následne podľa úspechov a neúspechov pri investovaní bayesovsky upravuje svoje presvedčenie. Payzan-LeNestour (2012) (pozri sekciu (3.3)) vo svojej behaviorálnej štúdii ukázali, že takéto spojenie "multi-armed restless bandits" s Bayesovským aktualizovaním dobre opisuje reálne finančné situácie.

Presvedčenia o pravdepodobnosti úspechu, prislúchajúce stavom v učiacom sa modeli sú reálne hodnoty, a preto nevieme odvodiť uzavretý Whittlov index, ale musíme jeho hodnotu simulovať. Použijeme algoritmus z Villar (2012), ktorý upravíme pre finančné aplikácie. Výpovednú silu tohto indexu overíme pomocou obsiahlej simulačnej štúdie. Vyskúšali sme viac ako sto rôznych scenárov. V práci uvádzame len reprezentatívnu vzorku dosiahnutých výsledkov. Simulačná štúdia je rozdelená do troch častí. V prvej ukazujeme rozdiely medzi jednotlivými porovnávacími pravidlami. V druhej poukazujeme na zmeny dosiahnuté pre rôzné diskontné faktory a v poslednej časti skúmame ako ovplyvňuje zmena parametrov dosiahnutý výsledok.

Simulácie ukazujú, že naše pravidlo má celkom dobrú úspešnosť. Na druhej sme si vedomí, že je ťažké porovnať naše pravidlo s reálne používanými stratégiami, keďže tie si investori držia v tajnosti. Myslíme si, že v budúcnosti by sa bolo dobré zaoberať zakomponovaním variancie nášho presvedčenia do indexu. Ako zaujímavé sa ukazuje aj preformulovanie modelu tak, aby sa výsledok dal priradiť k tabuľkovým hodnotám Gittinsovho indexu (index pre klasických banditov), čo by nám umožnilo porovnať ho so Sorensenovými výsledkami.

Bibliography

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics*, 16(3/4):221–229.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Blondel, V. and Tsitsiklis, J. N. (2000). A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274.

Bradt, R. N., Johnson, S. M., and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27(4):1060–1074.

Bruguier, Antoine J., S. R. Q. and Bossaerts, P. (2010). Exploring the nature of "trader intuition". *The Journal of Finance*, 65:1703–1723.

Crawford, G. and Shum, M. (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4):1137–1173.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177.

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. J. Wiley & Sons, New York.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J., editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam.

Guignard, M. (2003). Lagrangean relaxation. *TOP*, 11(2):151–228.

Hernández-Lerma, O. and Hoyos-Reyes, L. (2001). A multiobjective control approach to priority queues. *Mathematical Methods of Operations Research*, 53:265–277.

Hochberg, Y. V., Ljungqvist, A., and Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *Journal of Finance*, 62:251–301.

Jacko, P. (2009a). Adaptive greedy rules for dynamic and stochastic resource capacity allocation problems. *Medium for Econometric Applications*, 17(4):10–16. Available online at http://www.met-online.nl. Invited paper.

Jacko, P. (2009b). *Marginal Productivity Index Policies for Dynamic Priority Allocation in Restless Bandit Models*. PhD thesis, Universidad Carlos III de Madrid. http://e-archivo.uc3m.es/bitstream/10016/5357/1/tesis_jacko_peter.pdf.

Jacko, P. (2010a). *Dynamic Priority Allocation in Restless Bandit Models*. Lambert Academic Publishing. Invited book.

Jacko, P. (2010b). Restless bandits approach to the job scheduling problem and its extensions. In Piunovskiy, A. B., editor, *Modern Trends in Controlled Stochastic Processes: Theory and Applications*, pages 248–267. Luniver Press, United Kingdom.

Jacko, P. (2011). Value of information in optimal flow-level scheduling of users with markovian time-varying channels. *Performance Evaluation*, 68(11):1022–1036.

Kaplan, S. and Stromber, P. (2004). Characteristics, contracts, and actions: Evidence from venture capitalist analyses. *The Journal of Finance*, 59(5):2177–2210.

Michael L. Littman, A. R. C. and Kaelbling, L. P. (1995). Efficient dynamic-programming updates in partially observable markov decision processes. Technical report, Brown University.

Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98.

Niño-Mora, J. (2002). Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Mathematical Programming, Series A*, 93(3):361–413.

Niño-Mora, J. (2006). Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock $M/G/1$ queues. *Mathematics of Operations Research*, 31(1):50–84.

Niño-Mora, J. (2007). Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15(2):161–198.

Niño-Mora, J. (2010). Multi-armed restless bandits, index policies, and dynamic priority allocation. *Boletín de Estadística e Investigación Operativa*, 26(2):124–133.

Novak, V. (2011). Index policies for dynamic and stochastic problems.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1987). The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3):441–450.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queueing network. *Mathematics of Operations Research*, 24(2):293–305.

Pastor, L. and Veronesi, P. (2009). Learning in financial markets. *Chicago Booth School of Business Research Paper*, No. 08-28.

Payzan-LeNestour, E. (2012). Learning to choose the right investment in an unstable world: Experimental evidence based on the bandit problem. *Swiss Finance Institute Research Paper*, (10-28).

Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Quindlen, R. (2000). *Confessions of a venture capitalist : inside the high-stakes world of start-up financing*. New York : Warner Books.

Robbins, H. (1952a). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535.

Robbins, H. (1952b). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(527-535).

Sorensen, M. (2008). Learning by investing: Evidence from venture capital. *AFA 2008 New Orleans Meetings Paper*.

Stidham, S. J. (2002). Analysis, design, and control of queueing systems. *Operations Research*, 50(1):197–216.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:275–294.

Villar, S. S. (2012). *Restless Bandit Index Policies for Dynamic Sensor Scheduling Optimization*. PhD thesis, Universidad Carlos III de Madrid.

Visweswaran, V. (2009). Decomposition techniques for MILP: Lagrangian relaxation. In Floudas, C. and Pardalos, P., editors, *Encyclopedia of Optimization*. Springer, New York, 2nd edition edition.

Weber, R. (1992). On the Gittins index for multiarmed bandits. *Annals of Applied Probability*, 2(4):1024–1033.

Weber, R. and Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society, Series B*, 42(2):143–149.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *A Celebration of Applied Probability, J. Gani (Ed.), Journal of Applied Probability*, 25A:287–298.

Williams, J. L. (2007). *Information Theoretic Sensor Management*. PhD thesis, Massachusetts Institute of Technology.

Appendix

A. 1 Indexability of the multi-armed restless bandits

In this section we want to show a brief introduction of indexability for the multi-armed restless bandits. We follow the survey Niño-Mora (2007). In the following we use the notation described in the chapter 2 for the model 1 (see section (2.1.2)). Thus we focus on a discrete-time *single restless bandit model* (we want to show indexability for the decomposed single-company subproblem (equation (2.8))), with a finite state space $\mathcal{N}$, one-period rewards R_n^a and one-period state-transition probability $p_{n,m}^a$ if the company is in state n and after action a is chosen it evolves into state m. We consider the discounted case with factor $0 < \beta < 1$. The investor invests in the company under a policy π, drawn from the class Π of history-dependent policies. We denote by $X(t)$ and $a(t)$ the company state and action processes, respectively.

A policy π can be evaluated by two measures. The first one correspond to the *reward measure* and the second correspond to the *work measure*.

$$f_\tau^\pi := \mathbb{E}_\tau^\pi \left[\sum_{t=\tau}^{\infty} R_{X(t)}^{a(t)} \beta^t \right], \tag{A.1}$$

$$g_\tau^\pi := \mathbb{E}_\tau^\pi \left[\sum_{t=\tau}^{\infty} W_{X(t)}^{a(t)} \beta^t \right], \tag{A.2}$$

The reward measure (A.1) is equal to the expected total reward earned over an infinite horizon starting in a time instant τ. The work measure (A.2) describes the total discounted associated work (resource) expenditure. Usually, as in our case, we assume that $W_n^1 > W_n^0 \geq 0$ if the company is in the state $n \in X(t)$.

It is further important to identify states in which the work consumption and dynamics is identical. We denote such a set as $\mathcal{N}^{\{0\}}$ and call such states *uncontrollable*

$$\mathcal{N}^{\{0\}} := \{n \in \mathcal{N} : W_n^0 = W_n^1 \text{ and } p_{n,m}^0 = p_{n,m}^1, m \in \mathcal{N}\}.$$

We call the remaining states $\mathcal{N}^{\{0,1\}} := \mathcal{N}\backslash\mathcal{N}^{\{0\}}$ *controllable* and we denote by $i \geq 1$ the number of such states. Intuitively, the investor does not invest in uncontrollable states.

By drawing at random the initial state according to an arbitrary positive mass function $p_n > 0$, we obtain $f^\pi := \sum_{n\in N} p_n f_n^\pi$ and $g^\pi := \sum_{n\in N} p_n g_n^\pi$. Then the assumption that for work we pay ν-wage rate lead us to ν-wage problem

$$\max_{\pi\in\Pi} f^\pi - \nu g^\pi \tag{A.3}$$

where the goal is to find an admissible investment policy that maximize the value of rewards earned minus costs incurred. We use (A.3) as a *calibrating problem* aimed at measuring the marginal work at each company/industry state. The problem (A.3) is a finite-state and action discounted MDP, it is ensured existence of an optimal policy which is stationary deterministic and independent of initial state (Puterman (2005)). Each such a policy can be represented by its *active set* $\mathcal{S} \subseteq \mathcal{N}^{\{0,1\}}$. The active set is the subset of states where it is prescribed to invest in company. Therefore we can write $f^{\mathcal{S}}$ and $g^{\mathcal{S}}$ for $\mathcal{S}$-active policy. The problem (A.3) is reduced to the *combinatorial optimization problem*, where the goal is to find an optimal active set in the family of all subsets $2^{\mathcal{N}^{\{0,1\}}}$ of $\mathcal{N}^{\{0,1\}}$.

$$\max_{\mathcal{S}\in 2^{\mathcal{N}^{\{0,1\}}}} f^{\mathcal{S}} - \nu g^{\mathcal{S}} \tag{A.4}$$

Let denote $\widehat{V}_n$ the optimal value of (A.3) starting in state n. Then for every value ν, the optimal policies are prescribed by the Bellman equations unique solution

$$\widehat{V}_n(\nu) = \max_{a\in\{0,1\}} R_n^a - \nu W_n^a + \beta \sum_{m\in\mathcal{N}} p_{n,m}^a \widehat{V}_m(\nu), n \in \mathcal{N}, \tag{A.5}$$

So, *minimal optimal active set* $\widehat{\mathcal{S}} \subseteq \mathcal{N}^{\{0,1\}}$ for (A.3) exists and it is characterized in terms of (A.5) by

$$\widehat{\mathcal{S}}(\nu) := \left\{ n \in \mathcal{N}^{\{0,1\}} : R_n^1 - \nu W_n^1 + \beta \sum_{m\in\mathcal{N}} p_{n,m}^1 \widehat{V}_m(\nu) > R_n^0 - \nu W_n^0 + \beta \sum_{m\in\mathcal{N}} p_{n,m}^0 \widehat{V}_m(\nu) \right\}.$$

If in the model the active sets $\widehat{\mathcal{S}}(\nu)$ is expanding monotonically from the empty set $\emptyset$ to the full controllable state space $\mathcal{N}^{\{0,1\}}$ with the decreasing ν from ∞ to $-\infty$, then we can connect each controllable state n with a critical value $\widehat{\nu}_n$ below which n enters $\widehat{\mathcal{S}}(\nu)$.

Definition A. 1.1. *(Indexability) We say that company is indexable if there exists an index $\widehat{\nu}_n \in \mathbb{R}$ for $n \in \mathcal{N}^{\{0,1\}}$ such that*

$$\widehat{\mathcal{S}}(\nu) = \left\{ n \in \mathcal{N}^{\{0,1\}} : \widehat{\nu}_n > \nu \right\}, \nu \in \mathbb{R}.$$

Then $\widehat{\nu}_n$ is the company's marginal productivity (MP) index .

The set of MP indices $\widehat{\nu}_n$ for all $n \in N$ (if they exist) defines an optimal MP index policy: "Work if and only if the MP index of the current state is greater than the wage parameter ν."

Note that as we described in section (1.5) the MP index is a generalization of other indices.

A. 2 Work-reward view of indexability

Following Niño-Mora (2002, 2006) we introduce work-reward approach to indexability, which is deeply connected with the multi-objective optimization (see Hernández-Lerma and Hoyos-Reyes (2001)). Imagine the region of work-reward performance points in the plane under all admissible policies

$$\mathbb{H} := \{(g^\pi, f^\pi) : \pi \in \Pi\}.$$

We refer to this region as *achievable work-reward performance region.* This region is given by the convex hull of the finite collection of the performance points $(g^{\mathcal{S}}, f^{\mathcal{S}})$, where $\mathcal{S}$ represents the active sets of stationary deterministic policies which generate the performance points.

$$\mathbb{H} = conv\left(\left\{(g^{\mathcal{S}}, f^{\mathcal{S}}) : \mathcal{S} \in 2^{\mathcal{N}^{\{0,1\}}}\right\}\right).$$

The upper boundary $\bar{\partial}\mathbb{H}$ is defined by $(g, f) \in \mathbb{H}$ if and only if $f^\pi \leq f$ for every $\pi \in \Pi$ such that $g^\pi = g$.

The company is indexable iff $\bar{\partial}\mathbb{H}$ is characterized by a *nested active-set family*

$$\mathcal{F}_0 := \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_n\},$$

where $\mathcal{S}_0 := \emptyset$, $\mathcal{S}_n := \mathcal{N}^{\{0,1\}}$ and $\mathcal{S}_k := \{n_1, \ldots, n_k\}$ for $1 \leq k \leq i$ satisfy

$$g^{\mathcal{S}_0} < g^{\mathcal{S}_1} < \cdots < g^{\mathcal{S}_n},$$

and $n_1, \ldots, n_i$ is an ordering of the company's i controllable states. Thus, the MP index equals to

$$\widehat{\nu}_{n_k} = \frac{f^{\mathcal{S}_k} - f^{\mathcal{S}_{k-1}}}{g^{\mathcal{S}_k} - g^{\mathcal{S}_{k-1}}}, 1 \leq k \leq i. \tag{A.6}$$

If depicted in a plane with works on the x-axis and rewards on the y-axis, then the optimal policies to ((2.9) and (A.3)) lie on the upper boundary of such a region, since the parameter ν gives the slope of the supporting hyperplane (a line in this case) defining an optimum point (i.e., an optimal policy). In the figure (A.14) is an example of achievable work-reward performance region of an indexable company with three states $\mathcal{N} = \{1, 2, 3\}$ and in the figure (A.15) an example of a nonindexable company, where both examples are based on the examples from Niño-Mora (2007).

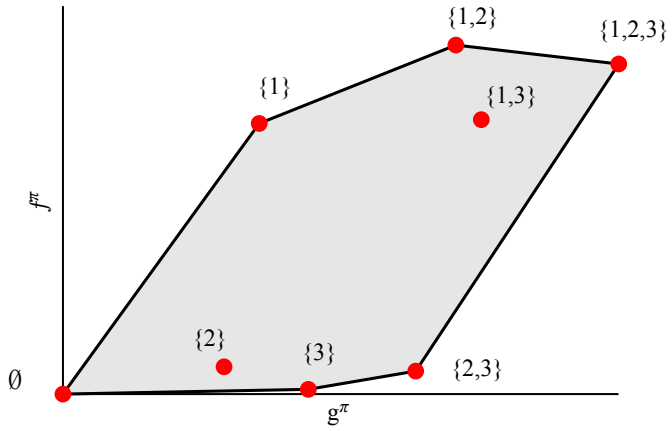

Figure A.14: Work-reward region for indexable company

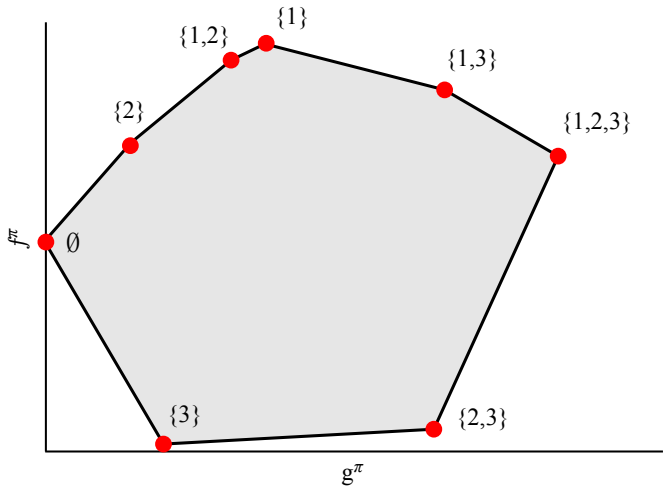

Figure A.15: Work-reward region for nonindexable company

A. 3 Adaptive-greedy algorithm

Decide about the indexability of the restless bandits in the figures above is easy task, because it could be done by visual inspection. On the other hand it is of a great practical interest to establish it analytically that some model is indexable under suitable parameter range. This task is generally far from trivial. Niño-Mora (2001, 2002) and Niño-Mora (2006) developed tractable sufficient conditions for indexability that are widely applicable, together with an index algorithm.

We denote by $\langle a, \mathcal{S} \rangle$ the policy that choose action $a \in \{0, 1\}$ in the initial instant and adopts the $\mathcal{S}$-active policy thereafter. Moreover, we define the *marginal work measure*

$$w_n^{\mathcal{S}} := g_n^{\langle 1,\mathcal{S}\rangle} - g_n^{\langle 0,\mathcal{S}\rangle} = W_n^1 - W_n^0 + \beta \sum_{m\in\mathcal{N}} \left(p_{n,m}^1 - p_{n,m}^0\right) g_m^{\mathcal{S}},$$

we also define the *marginal reward measure*

$$r_n^{\mathcal{S}} := f_n^{\langle 1,\mathcal{S}\rangle} - f_n^{\langle 0,\mathcal{S}\rangle} = R_n^1 - R_n^0 + \beta \sum_{m\in\mathcal{N}} \left(p_{n,m}^1 - p_{n,m}^0\right) f_m^{\mathcal{S}}.$$

and for the case when $w_n^{\mathcal{S}} \neq 0$, we define the *marginal productivity measure*

$$\nu_n^{\mathcal{S}} := \frac{r_n^{\mathcal{S}}}{w_n^{\mathcal{S}}}.$$

Usually we are not able to identify a priori the nested active-set family $\mathcal{F}_0$, unless the state space is linearly ordered. We can try to guess the structure of optimal policies for the particular model in the form $\mathcal{F} \subseteq 2^{\mathcal{N}^{\{0,1\}}}$ containing $\mathcal{F}_0$. Unfortunately, $\mathcal{F}$ is often much larger than $\mathcal{F}_0$. In the terminology of combinatorial optimization, $(\mathcal{N}^{\{0,1\}}, \mathcal{F})$ is a set system on ground set $\mathcal{N}^{\{0,1\}}$ having $\mathcal{F}$ as its family of feasible sets.

For developing an algorithm it must be satisfied that we can proceed from the empty set towards a given set $\mathcal{S} \in \mathcal{F}$ through successive single-state augmentations. Moreover, we have the symmetric requirement for reaching $\mathcal{S}$ through successive single-state removals from $\mathcal{N}^{\{0,1\}}$. Thus, we define the *inner boundary* of $\mathcal{S}$ relative to $\mathcal{F}$, for $\mathcal{S} \in \mathcal{F}$ by

$$\partial_{\mathcal{F}}^{in}\mathcal{S} := \{n \in \mathcal{S} : \mathcal{S} \setminus \{n\} \in \mathcal{F}\},$$

and we define the *outer boundary* of $\mathcal{S}$ relative to $\mathcal{F}$ by

$$\partial_{\mathcal{F}}^{out}\mathcal{S} := \left\{n \in \mathcal{N}^{\{0,1\}} \setminus \mathcal{S} : \mathcal{S} \cup \{n\} \in \mathcal{F}\right\},$$

so we can set the following assumption.

Assumption A. 3.1. *Set system* $(\mathcal{N}^{\{0,1\}}, \mathcal{F})$ *satisfies the following conditions:*

1. $\emptyset, \mathcal{N}^{\{0,1\}} \in \mathcal{F}$
2. *For* $\emptyset \neq \mathcal{S} \in \mathcal{F}, \partial^{out}_{\mathcal{F}} \mathcal{S} \neq \emptyset$
3. *For* $\mathcal{N}^{\{0,1\}} \neq \mathcal{S} \in \mathcal{F}, \partial^{in}_{\mathcal{F}} \mathcal{S} \neq \emptyset$

Now we can define the *adaptive-greedy algorithm* $AG_{\mathcal{F}}$ in the following way.

Algorithm $AG_{\mathcal{F}}$
output: $\{n_k, \widehat{\nu}_{n_k}\}_{k=1}^{i}$
$\mathcal{S}_0 := \emptyset$
for $k := 1$ *to* i **do**
  pick $n_k \in \arg\max \left\{ \nu_n^{\mathcal{S}_{k-1}} : n \in \partial^{out}_{\mathcal{F}} \mathcal{S}_{k-1} \right\}$;
  $\widehat{\nu}_{n_k} := \nu_{n_k}^{\mathcal{S}_{k-1}}; \mathcal{S}_k := \mathcal{S}_{k-1} \cup \{n_k\}$;
end

Algorithm 2: $AG_{\mathcal{F}} - algorithm$

From the geometric view point aim of this algorithm is to traverse the upper boundary of the achievable work-reward performance region, building up the succesive active sets $\mathcal{S}_k$ forming the nested family $\mathcal{F}_0$ that determines such a boundary. The algorithm traverse the upper boundary from left to right and it is a top-down algorithm, so the successive index values $\widehat{\nu}_{n_k}$ or slopes in such a frontier are computed in nonincreasing order.