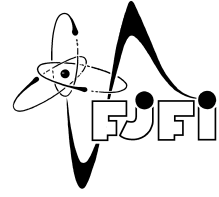




CZECH TECHNICAL UNIVERSITY IN PRAGUE  
Faculty of Nuclear Sciences and Physical Engineering



# **Dynamic Fully Probabilistic Decision Making with Stopping**

## **Dynamické plně pravděpodobnostní rozhodování se zastavováním**

Master's Thesis

Author: **Bc. Daniel Karlík**  
Supervisor: **Ing. Miroslav Kárný, DrSc.**  
Academic year: 2023/2024

## I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Karlík** Jméno: **Daniel** Osobní číslo: **476591**  
Fakulta/ústav: **Fakulta jaderná a fyzikálně inženýrská**  
Zadávací katedra/ústav: **Katedra matematiky**  
Studijní program: **Aplikované matematicko-stochastické metody**

## II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

**Dynamické plně pravděpodobnostní rozhodování se zastavováním**

Název diplomové práce anglicky:

**Dynamic Fully Probabilistic Decision Making with Stopping**

Pokyny pro vypracování:

- 1) Prohlubte své znalosti plně pravděpodobnostního návrhu (PPN) rozhodovacích strategií na případy se smíšenými (diskrétními a spojitými daty) a vnořte úlohy s (případně částečným) zastavováním do tohoto rámce, [1], [9].
- 2) Popište typické úlohy dynamického rozhodování, u nichž je zastavování významné, viz např. [2], [3], [4] a případně další.
- 3) Rozšířte praktickou použitelnost Vašeho obecného řešení o algoritmickou podporu kvantitativní volby uživatelských vstupů do PPN se zastavováním, např. ve směrech plynoucích z [5], [6].
- 4) Algoritmizujte své řešení obohacené o výsledky z bodu 3.) pro implementovatelné případy dynamických systémů [7].
- 5) Ilustrujte vlastnosti Vašeho řešení dle bodu 4.) simulačně [8].

Seznam doporučené literatury:

- [1] M. Kárný, T. V. Guy, Fully probabilistic control design. *Systems & Control Letters* 55(4), 2006, 259–265.
- [2] A. Wald, *Sequential Analysis*. Dover Publications, 2013.
- [3] T. S. Ferguson, Who solve the secretary problem?. *Statistical Science* 4(3), 1989, 282–296.
- [4] M. Kárný, Fully Probabilistic Design of Strategies with Estimator. *Automatica* 141, 2022, 110269.
- [5] V. Peterka, Bayesian Approach to System Identification. In 'Trends & Progress in System Identification', Pergamon Press, 1981, 239–304.
- [6] M. Kárný, T. Siváková, Agent's Feedback in Preference Elicitation. In 'IUCC-CSS 2021', 2021, 421–429.
- [7] M. Puterman, *Markov Decision Processes*. John Wiley & Sons, 1994.
- [8] M. Virius, *Metoda Monte Carlo*. Vydavatelství ČVUT, 2010.
- [9] D. Karlík, Dynamic decision making with stopping. Research project Dept. Mathematics, FNSPE CTU, 2023.

Jméno a pracoviště vedoucí(ho) diplomové práce:

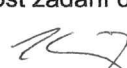
**Ing. Miroslav Kárný, DrSc. ÚTIA AV ČR Praha**


Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

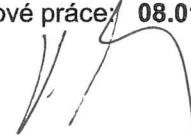
Datum zadání diplomové práce: **10.03.2023**

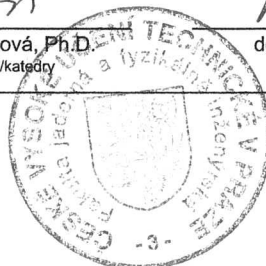
Termín odevzdání diplomové práce: **08.01.2024**

Platnost zadání diplomové práce: **10.03.2025**

  
Ing. Miroslav Kárný, DrSc.  
podpis vedoucí(ho) práce

  
prof. Ing. Zuzana Masáková, Ph.D.  
podpis vedoucí(ho) ústavu/katedry

  
doc. Ing. Václav Čuba, Ph.D.  
podpis děkana(ky)



### III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

3.4.2023

Datum převzetí zadání

Karlík

Podpis studenta

*Acknowledgment:*

I express my sincere gratitude to my supervisor, Ing. Miroslav Kárný, DrSc., for his invaluable guidance, patience, and unwavering support not only during the completion of this Master's Thesis but throughout my entire academic journey. I extend my appreciation to the members of the Department of Adaptive Systems in UTIA for their constructive feedback during the development of this work. A special acknowledgment goes to MgA. Karolína Stará for her steadfast support whenever it was needed. Additionally, I would like to thank Ing. Tatiana Valentine Guy, PhD., for the support of this project with the grant EU-COST Action CA21169.

*Author's declaration:*

I declare that this Master's Thesis is entirely my own work and I have listed all the used sources in the bibliography.

Prague, January 7, 2024

Bc. Daniel Karlík

*Název práce:*

## **Dynamické plně pravděpodobnostní rozhodování se zastavováním**

*Autor:* Bc. Daniel Karlík

*Program:* Aplikované matematicko-stochastické metody

*Druh práce:* Diplomová práce

*Vedoucí práce:* Ing. Miroslav Kárný, DrSc., ÚTIA AV ČR Praha

*Abstrakt:* Tato práce přispívá k řešení problematiky v oblasti rozhodování, přesněji v oblasti dynamického rozhodování, kde je důležité provádět rozhodování s ohledem na čas a vývoj problému. Cílem této práce bylo navrhnout řešení jež by vedlo k získání optimálního zastavovacího pravidla. K řešení rozhodovacího procesu jsme použili teorie plně pravděpodobnostního návrhu (PPN). PPN představuje rozšíření tradiční rozhodovací metody markovských rozhodovacích procesů. PPN řeší a modeluje vývoj stavů v uzavřené rozhodovací smyčce pomocí ideální distribuce chování, která reprezentuje agentovy preference. Optimální rozhodovací pravidlo v PPN je hledáno pomocí minimalizace Kullback-Leiblerovi divergence mezi distribucí chování modelu a jejím ideálem. Metoda zjišťování preferencí v PPN slouží k převodu preferencí na ideální distribuce chování agenta. V této práci jsme navrhli rozšíření PPN o zastavování a poskytli jeho řešení. Dále jsme navrhli způsob využití zjišťování preferencí v řešení PPN se zastavováním. Nakonec jsme pomocí simulovaných experimentů a bayesovského odhadování parametrů ověřili kvalitu a rychlost námi navrženého řešení.

*Klíčová slova:* bayesovské odhadování parametrů, dynamické rozhodování, Kullback-Leiblerova divergence, markovské rozhodovací procesy, plně pravděpodobnostní návrh, zastavovací pravidlo, zjišťování preferencí

*Title:*

## **Dynamic Fully Probabilistic Decision Making with Stopping**

*Author:* Bc. Daniel Karlík

*Abstract:* This work contributes to the problem in the area of decision making, more specifically in the area of dynamic decision making, where it is important to make decisions with respect to time and the evolution of the problem. The aim of this work was to propose a solution that would lead to obtain an optimal stopping rule. We used the theory of Fully probabilistic design (FPD) to solve the decision making processes. FPD is an extension of the well known Markov decision processes. FPD solves and models the evolution of states in a closed decision loop using an ideal probability distribution (pd) of behavior that represents the agent's preferences. The optimal decision rule in FPD is evaluated by minimizing the Kullback-Leibler divergence between the model's pd of behavior and its ideal. The preference elicitation (PE) method in FPD is used to convert the preferences into the agent's ideal pd of behaviors. In this work, we propose an extension of FPD to incorporate stopping and provide its solution. We also proposed an approach to exploit PE in the solution of FPD with stopping. Finally, we verified the quality and speed of our proposed solution using simulated experiments and Bayesian parameter estimation.

*Key words:* Bayesian parameter estimation, dynamic decision making, fully probabilistic design, Kullback-Leibler divergence, Markov decision processes, preference elicitation, stopping rule

# Contents

<b>Introduction</b>	<b>7</b>
<b>1 Preliminaries</b>	<b>9</b>
1.1 Probability Theory and Notation . . . . .	9
1.2 Markov Decision Process . . . . .	10
1.3 Fully Probabilistic Design of Decision Strategies . . . . .	12
1.4 Preference Elicitation . . . . .	13
<b>2 FPD with Stopping</b>	<b>16</b>
2.1 Design of the Extended Behaviour . . . . .	16
2.2 Design of the Extended Ideal Probability . . . . .	18
2.3 Evaluation of the Optimal Policy using FPD with Stopping . . . . .	19
<b>3 Preference Elicitation in FPD with Stopping</b>	<b>22</b>
3.1 Algorithm for PE . . . . .	22
3.2 Incorporation of Preference Elicitation into FPD with Stopping . . . . .	23
3.3 Bayesian Parameter Estimation . . . . .	24
<b>4 Simulations</b>	<b>26</b>
4.1 Design of the Simulation . . . . .	26
4.2 Performance Metrics . . . . .	27
4.3 Input Settings . . . . .	27
4.4 First Experiment . . . . .	27
4.5 Second Experiment . . . . .	29
4.6 Third Experiment . . . . .	31
4.7 Discussion . . . . .	37
<b>5 Conclusion</b>	<b>41</b>

# Introduction

All living beings share a fundamental characteristic: the necessity to make decisions. While some decisions are instinctive and occur automatically, others are complex and require significant time and attention, see [7]. For example, consider a wild animal deciding whether to take the quickest or safest path to water – a decision that balances urgency against risk. This scenario offers just an insight into the omnipresence of decision making (DM) in almost every aspect of life.

Recognizing the importance of DM, this work emphasizes its role as a crucial and comprehensive field of study. DM integrates knowledge from diverse disciplines to develop effective tools for tackling complex problems. Effective decision making should be grounded in relevant knowledge, despite the common human tendency to let emotions and extraneous information cloud the judgment. For more information on rationality in non-rational DM, see [10], as an interesting but not yet fully explored area of study.

We model DM processes using a closed loop comprising a system and an agent. The system represents the world segment the agent aims to understand and influence. The agent observes this system, deciding on actions which he use to influence the system and achieve preferred outcomes from the system. We refer to these outcomes as states. The agent generates actions, whereas the system generates states.

Our work builds upon the Markov decision process (MDP) formalism [14], a mathematical framework for describing this closed-loop interactions. The goal of a rational agent within this framework is to select an optimal policy, a set of rules guiding its actions. MDP theory aids in finding the most effective policy, incorporating the agent's preferences and all relevant information, such as observations of the system, expert insights, and other external knowledge.

The conditions under which agents seek optimal policies in the closed loop vary widely. Typical difference is in the knowledge of the system model. The models of the system may be known or unknown, and in the latter case, its parameters can be estimated using, for instance, Bayesian parameter estimation [2].

This work focuses on DM under time constraints, a practical consideration given that the optimizing decision making incurs additional costs and utilizes computational resources before resolution. This problem, of when to stop the process of DM and make the decision or use the policy, is DM with a stopping rule. This dynamic DM process is a part of the extensive field of sequential DM. For further details, refer to Wald's seminal work on sequential analysis, see [19].

One of the most renowned examples involving the stopping rule is the Secretary problem. This problem likely emerged in the late 1950s, although similar concepts have been explored for centuries, such as by A. Cayley in 1875 and possibly by J. Kepler in the early 17th century. For a comprehensive history of this problem, refer to [5]. A detailed and exhaustive solution of this problem is presented in [12].

Another significant application of the stopping rule is the diagnosis-versus-treatment dilemma in medicine. A doctor must diagnose a patient's disease in a timely manner to ensure that it is not too late for effective treatment. For further details, refer to [20].

The list of examples requiring a stopping rule is extensive, including decisions like determining the end of an experiment or ceasing the estimation of certain parameters.

A significant challenge in DM, particularly under time pressure, is the enhanced tendency towards irrational DM.

The aim of this work is to propose a robust methodology for optimizing the stopping rule, employing Fully Probabilistic Design of decision strategies (FPD) theory. FPD, a relatively recent concept, is yet to be fully explored in the context of DM with stopping rules. Unlike MDP, which utilizes a loss function to evaluate the optimal policy, FPD probabilistically quantifies agent's aims and preferences. In FPD the agent's preferences are represented in the so called ideal probability distribution (pd). High values are assigned to preferred states and actions, low to unpreferred ones. The optimal policy is evaluated as an argument minimizing the Kullback-Leibler divergence of the real pd and ideal pd. For a closer look on FPD theory and its formulation see [9] and [10]. This work builds on work [6], and seeks to complete its unfinished solution, dealing with stopping in FPD. Presently, there is no general solution for stopping in FPD or for a partial stopping in dynamic DM.

The agent need not be completely independent. It may have to make decisions in order to fulfill the preferences of a third party, which we call the user. The relationship between the agent and the user can be thought of as a relationship between the requestor of a task and its handler. However, in this work, we will only consider cases where the agent makes decisions with the aim of satisfying the specified preferences. Of course, it may be the case that the agent and the user are the same person, but in this work we distinguish these two roles. It allows us to cope with the fact the user expresses preference incompletely and not in the form of ideal pd. The needed transformation how can the agent reflects given aims and preferences into DM is called preference elicitation (PE). Use of the PE in FPD provides a transformation of user's preferences into non-empty set of prospective ideal pd and then making a choice of the optimal ideal pd, see [10]. This approach is much easier with FPD mathematical framework than with MDP, because working with the utility function might be more difficult from the designing perspective: the deductive rules how to combine partial losses are missing, refer to [3], [4].

This work introduces an extension of FPD to incorporate DM with stopping. We have constructed a theoretical solution of FPD with stopping. Then we have combined this solution with PE and proposed the way how to evaluated the optimal policy with stopping. The effectiveness of the proposed solution has been validated by experimental testing.

Chapter 1 recalls the theory necessary for this work, covering the basics of probability theory, MDP, FPD and PE. Chapter 2 elaborates on the extension of FPD to include stopping and outlines the formulation of newly established ideal pds, which include the stopping. This chapter results in the solution of FPD with stopping. Chapter 3 explores how PE can be incorporated into FPD with stopping. In Chapter 4, we introduce the simulated system used for testing, the studied metrics and various examples with different settings. Here, our proposed method is directly compared with the standard method that do not involve stopping. The results of these experiments are presented and briefly discussed. Finally, Chapter 5 offers concluding thoughts and ideas about possible future research related to this work.



# Chapter 1

## Preliminaries

In this chapter, we introduce the theory to the reader to the extent that is sufficient for the purposes of this work. We will also introduce the notation that will be used throughout this work.

### 1.1 Probability Theory and Notation

In this section, we present only basic concepts of probability theory and auxiliary functions used in the following sections and chapters. For the purpose of this work, we have chosen to adopt an engineering perspective on the topic of probability theory. More deep and exhaustive view on probability field of study can be found in [15].

**Remark 1.** Throughout,  $p(\cdot), m(\cdot), \pi(\cdot), c(\cdot)$  always represent probability distributions (pds), where  $p(x)$  and  $p(a)$  are two different pds which are distinguished by labels of the different random variables  $x$  and  $a$ .

In the text we also use different font for other functions, e.g.  $d(\cdot), h(\cdot)$ , we tried to distinguish these functions, to not be mistaken with pds.

Notation	Meaning
$\mathcal{X}$	Set of values of the variable X
$ \mathcal{X} $	Cardinality of the set $\mathcal{X}$
$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathcal{T}$	Set of time moments
$\mathcal{S}$	Set of states
$\mathcal{A}$	Set of actions
$\equiv$	Definition by assignment
$\pi \equiv (\pi(a_t s_{t-1}))_{t=1}^{ \mathcal{T} }$	Policy, sequence of pds
$\Pi$	Set of policies
$(m(s_t a_t, s_{t-1}))_{t=1}^{ \mathcal{T} }$	System model
$\propto$	Proportionality

Table 1.1: The notation that is used in the work.

**Definition 1.** A marginal probability distribution (pd) is defined using a joint probability distribution  $p(a, b)$  by the following relation

$$p(a) = \int_{\mathcal{B}} p(a, b) db,$$

where  $\mathcal{B}$  denotes the set of all possible values of  $b$ .

**Definition 2.** The conditional pd of the variable  $a$ , conditioned on the occurrence of  $b$ , is defined by

$$p(a|b) = \frac{p(a, b)}{p(b)}, \quad p(b) > 0.$$

**Definition 3.** The independence of the variable  $a$  from the variable  $b$  is defined as follows

$$p(a, b) = p(a)p(b)$$

or equivalently

$$p(a|b) = p(a).$$

**Theorem 1.** The following formula is known as the chain rule

$$p(x_n, x_{n-1}, \dots, x_1) \equiv p(x_n, x_{n-1}, \dots, x_1|x_0) = \prod_{k=1}^n p(x_k|x_{k-1}, \dots, x_0). \quad (1.1)$$

*Proof.* The chain rule can be obtained by a repeated use of definition 2. □

**Remark 2.** In this work the knowledge of  $x_0$  is considered implicitly. This knowledge influences all other related relations and functions.

**Definition 4.** *Kronecker delta function* is denoted as  $\delta(\bullet, \bullet)$  and is defined as follows

$$\delta(x, y) \equiv \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases} \quad (1.2)$$

**Definition 5.** *The indicator function of a subset  $\tilde{\mathcal{X}}$  of  $\mathcal{X}$*  is defined as

$$1_{\tilde{\mathcal{X}}}(x) \equiv \begin{cases} 1 & \text{if } x \in \tilde{\mathcal{X}}, \\ 0 & \text{if } x \notin \tilde{\mathcal{X}}. \end{cases} \quad (1.3)$$

**Definition 6.** Let  $k \geq 1$  be a real number. The  $k$ -norm of row vector  $\mathbf{x} \equiv (x_1, \dots, x_n)$  with real-valued entries  $x_i$  for  $i = 1, \dots, n$ , is

$$\|\mathbf{x}\|_k \equiv \left( \sum_{i=1}^n |x_i|^k \right)^{1/k}, \quad n \in \mathbb{N}. \quad (1.4)$$

**Definition 7.** Suppose that  $f : \mathcal{X} \rightarrow \mathbb{R}^+$  is a non-negative valued function whose domain is an arbitrary set  $\mathcal{X}$ . The *support* of the function  $f$  is defined as:

$$\text{supp}(f) \equiv \{x \in \mathcal{X} | f(x) \neq 0\}. \quad (1.5)$$

## 1.2 Markov Decision Process

In this section we focus on presenting *closed loop* description, already mentioned in Introduction, but now in a more formal way and with the proper terminology used across this work.

The key structure studied in this work is called *closed loop*. A closed loop consists of a *system* and an *agent*. The *system* generates states according to some, possibly unknown, random rules. These states are observed by an *agent*, which generates actions that enter the system and possibly influence the closed loop. The system can represent a part of the world or a general problem we want to study, while the agent represents a decision maker, who wants to study the unknown system or influence it.

Our formal description exploits the next definition.

**Definition 8.** A Discrete Markov Decision Process is defined as a 5-tuple  $\{\mathcal{T}, \mathcal{S}, \mathcal{A}, m, l\}$

- $\mathcal{T} \equiv \{1, \dots, |\mathcal{T}|\}$ , is the set of time labels and  $|\mathcal{T}| \in \mathbb{N}$  is a fixed *horizon*.
- $\mathcal{S} \equiv \{s^i\}_{i=0}^{|\mathcal{S}|}$ , where  $|\mathcal{S}| \in \mathbb{N}$  is the *finite state space*, with given values  $s^i$ .
- $\mathcal{A} \equiv \{a^i\}_{i=0}^{|\mathcal{A}|}$ , where  $|\mathcal{A}| \in \mathbb{N}$  is the *finite action space*, with given values  $a^i$ .
- $s_t \in \mathcal{S}$  is the *state* obtained at time  $t \in \mathcal{T}$ ;  $a_t \in \mathcal{A}$  is the *action* obtained at time  $t \in \mathcal{T}$ .
- $m(s_t|a_t, s_{t-1})$  is the *system model*, which is the probability that the state  $s_{t-1}$  at time  $t-1$  and action  $a_t$  at time  $t$ , lead to the state  $s_t$  at time  $t$ .
- $l(s_t, a_t, s_{t-1})$  is the *loss function*, i.e. the function  $l : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ .

With respect to the DM domain, we use the following objects related to MDP.

**Definition 9.** As a *decision rule* we understand the probability  $\pi(a_t|s_{t-1})$ . The sequence of decision rules from  $t = 1$  up to  $t = |\mathcal{T}|$  forms the *policy*.

**Definition 10.** The *total loss function* is defined by the formula  $L \equiv \sum_{t \in \mathcal{T}} l(s_t, a_t, s_{t-1})$ .

**Definition 11.** The definition of *expected loss* under the policy  $\pi$  is as follows

$$\mathbb{E}^\pi[l(s_t, a_t, s_{t-1})] \equiv \sum_{\substack{s_t, s_{t-1} \in \mathcal{S} \\ a_t \in \mathcal{A}}} l(s_t, a_t, s_{t-1})m(s_t|a_t, s_{t-1})\pi(a_t|s_{t-1})c^\pi(s_{t-1}), \quad (1.6)$$

where  $c^\pi(s_{t-1})$  is marginal probability of the state  $s_{t-1}$ .

**Definition 12.** The *total expected loss* under the policy  $\pi$  is defined as

$$\mathbb{E}^\pi[L] \equiv \sum_{t \in \mathcal{T}} \sum_{\substack{s_t, s_{t-1} \in \mathcal{S} \\ a_t \in \mathcal{A}}} l(s_t, a_t, s_{t-1})p(s_t|a_t, s_{t-1})\pi(a_t|s_{t-1})c^\pi(s_{t-1}), \quad (1.7)$$

where the marginal probability  $c^\pi(s_{t-1})$  depends on the used policy.

**Definition 13.** The *value-function* of a policy  $\pi$  is defined as follows

$$u^\pi(s_{t-1}) \equiv \mathbb{E}^\pi \left[ \sum_{\substack{\tau \in \mathcal{T} \\ \tau \geq t}} l(s_\tau, a_\tau, s_{\tau-1}) \middle| s_{t-1} \right] \text{ for } \forall s_{t-1} \in \mathcal{S}, \quad (1.8)$$

where  $\mathbb{E}^\pi [l(s_\tau, a_\tau, s_{\tau-1})|s_{t-1}] \equiv \sum_{a_\tau \in \mathcal{A}} \sum_{s_\tau \in \mathcal{S}} l(s_\tau, a_\tau, s_{\tau-1})m(s_\tau|a_\tau, s_{\tau-1})\pi(a_\tau|s_{\tau-1})c^\pi(s_{\tau-1}|s_{t-1})$  for  $\tau \geq t$ .

**Definition 14.** The *optimal value-function* is defined by the following formula

$$u^{\pi^o}(s_{t-1}) \equiv \min_{\left\{ \begin{array}{l} \pi(a_\tau|s_{\tau-1}) \\ \tau \geq t \end{array} \right\}} \mathbb{E}^\pi \left[ \sum_{\substack{\tau \in \mathcal{T} \\ \tau \geq t}} l(s_\tau, a_\tau, s_{\tau-1}) \middle| s_{t-1} \right] \text{ for } \forall s_{t-1} \in \mathcal{S}. \quad (1.9)$$

**Definition 15.** The *optimal policy* is defined as

$$\pi^o \in \arg \min_{\pi \in \Pi} u^\pi(s_0). \quad (1.10)$$

The *optimal policy* can be found through the usage of the backward recursion, which is also called *dynamic programming*. Details about dynamic programming can be found in [16] and [19]. A more detailed discussion of the MDP is available in [14].

### 1.3 Fully Probabilistic Design of Decision Strategies

This section provides an overview of FPD theory to the extent that is sufficient for the purposes of this work. FPD represents an extension of Bayesian DM theory, effectively generalizing the MDP framework. For a more comprehensive explanation of FPD theory, readers can found in [9] and [10].

Below, we present exploitation of the introduced *closed loop* notation in the context of FPD theory. In this work, we restrict our focus to discrete valued states and actions.

**Definition 16.** *Behaviour of closed loop* up to the finite horizon  $|\mathcal{T}| \in \mathbb{N}$  is

$$b \equiv (s_{|\mathcal{T}|}, a_{|\mathcal{T}|}, s_{|\mathcal{T}|-1}, \dots, s_1, a_1) \in \mathcal{B}, s_0 \text{ fixed} \quad (1.11)$$

where  $s_t \in \mathcal{S}$  for  $\forall t \in \{0\} \cup \mathcal{T}$  and  $a_t \in \mathcal{A}$  for  $\forall t \in \mathcal{T}$  and  $\mathcal{B}$  denotes the *set of all behaviours*.

**Remark 3.** Thus, the behaviour  $b$  represents the sequence of consecutive states and actions in a closed loop for given  $s_0$

**Definition 17.** *Probability of behaviour in the closed loop* is

$$c^\pi(b) \equiv c^\pi(b|s_0) = \prod_{t \in \mathcal{T}} m(s_t|a_t, s_{t-1})\pi(a_t|s_{t-1}) \equiv p(b)\pi(b). \quad (1.12)$$

The transition probabilities  $m(s_t|a_t, s_{t-1})$ , modelling system in (1.12) are also referred in the literature as the *predictors* especially when they result from estimation.  $\pi(a_t|s_{t-1})$  is the *decision rule* see Definition 9.

**Definition 18.** *Ideal probability of behaviour of the closed loop* is defined as follows

$$c^i(b) \equiv c^\pi(b|s_0) = \prod_{t \in \mathcal{T}} m^i(s_t|a_t, s_{t-1})\pi^i(a_t|s_{t-1}) \equiv m^i(b)\pi^i(b) \quad (1.13)$$

and it describes the desired behaviours of a closed loop.  $m^i, \pi^i$  are ideal counterparts of the transition probabilities (models) and decision rules.

**Definition 19.** *Kullback-Leibler divergence* (KLD) of probabilities  $f(b), g(b)$  on  $b \in \mathcal{B}, |\mathcal{B}| < \infty$ , where  $g(b) > 0$  whenever  $f(b) > 0$ , is defined as

$$D(f||g) \equiv \sum_{b \in \mathcal{B}} f(b) \ln \frac{f(b)}{g(b)}. \quad (1.14)$$

**Remark 4.** Kullback-Leibler divergence, as defined in (1.14), is used and studied in various areas of mathematics, e.g. [17].

**Definition 20.** *Optimal policy in FPD* reads

$$\pi^o \in \arg \min_{\pi \in \Pi} D(c^\pi || c^i), \quad (1.15)$$

where  $c^\pi$  and  $c^i$  are defined as (1.12) and (1.13) respectively.

**Theorem 2.** KLD has the following additive form

$$D(c^\pi || c^i) = \sum_{t \in \mathcal{T}} \sum_{s_{t-1} \in \mathcal{S}} c^\pi(s_{t-1}) \sum_{\substack{s_t \in \mathcal{S} \\ a_t \in \mathcal{A}}} m(s_t|a_t, s_{t-1})\pi(a_t|s_{t-1}) |l^\pi(s_t, a_t, s_{t-1})|. \quad (1.16)$$

*Proof.* To see the complete proof of Theorem 2, refer to [6], page 21. □

**Definition 21.** Loss function in FPD is defined with respect to selected policy, see Definition 9

$$l^\pi(s_t, a_t, s_{t-1}) \equiv \ln \left( \frac{m(s_t|a_t, s_{t-1})\pi(a_t|s_{t-1})}{m^i(s_t|a_t, s_{t-1})\pi^i(a_t|s_{t-1})} \right). \quad (1.17)$$

**Remark 5.** FPD is specific by the dependence of the loss  $l^\pi$  on the policy  $\pi$ . This makes  $D(c^\pi \| c^i)$  nonlinear in  $\pi$  and the optimal policy randomized.

**Theorem 3.** (Solution of FPD) The optimal policy can be found using the following formulae, for  $t \in \mathcal{T}$ ,

$$\pi^o(a_t|s_{t-1}) = \pi^i(a_t|s_{t-1}) \frac{\exp[-d(a_t, s_{t-1})]}{h(s_{t-1})}, \quad (1.18)$$

where

$$h(s_{t-1}) \equiv \sum_{a_t \in \mathcal{A}} \pi^i(a_t|s_{t-1}) \exp[-d(a_t, s_{t-1})] \leq 1, \text{ for } t < |\mathcal{T}|, \quad (1.19)$$

while  $h(s_{|\mathcal{T}|}) = 1, \forall s_{|\mathcal{T}|} \in \mathcal{S}$  and

$$d(a_t, s_{t-1}) \equiv \sum_{s_t \in \mathcal{S}} m(s_t|a_t, s_{t-1}) \ln \left( \frac{m(s_t|a_t, s_{t-1})}{h(s_t)m^i(s_t|a_t, s_{t-1})} \right). \quad (1.20)$$

The solution is gained against the time course, starting at  $t = |\mathcal{T}|$ , using the backward recursion. Using decision rule, model on predictors and ideal models we can evaluate the optimal policy  $\pi^o$  in FPD and provide the minimum of

$$\min_{\pi \in \Pi} D(c^\pi \| c^i) = -\ln h(s_0) = u^{\pi^o}(s_0),$$

where the value function  $u^\pi$  is defined as (1.8).

*Proof.* The proof of Theorem 3 can be found in [9], where authors provides a detailed exposition in Chapter 3.  $\square$

## 1.4 Preference Elicitation

In this section, we aim to introduce the preference elicitation (PE) approach to the extent sufficient for this work. For a more detailed and in-depth insight into this field of study, refer to [11].

The reason for exploiting the preference elicitation approach is to construct the ideal pd  $c^i$  in (1.15), which reflects the preferences of the agent. Once we have  $c^i$ , optimal DM policy  $\pi^o$  can be evaluated as shown in (1.15). The use of PE within FPD leads towards the selection of the pd  $c^{i^o}$ , which represents the agent's incomplete preferences expressed in a non-probabilistic way. The agent's preferences define the set of ideal pds.

The incompleteness in the description of the preferences implies that the set  $C^i$  of ideal pds  $c^i$ , acting on  $\mathcal{B}$ , might contain a large number of pds, potentially even infinitely many. The set  $C^i$  can be empty as well, due to the inconsistencies in the agent's preferences, e.g. "I do not want to pay anything for heating, but I want it to be 25°C during winter".

In PE, is dealt with the non-empty set  $C^i$ , obtained by the use of reachable and reasonable preferences or by transforming the agent's preferences in a reasonable way to reflect his preferences while ensuring a non-empty set  $C^i$ .

The PE principle chooses the optimal ideal pd as follows:

$$c^{i^o} \in \arg \min_{c^i \in C^i} \left[ \min_{\pi \in \Pi} D(c^\pi \| c^i) \right]. \quad (1.21)$$

Use of this principle ensures that no other preferences or constraints are added to those expressed by the agent.

The minimizing over  $c^{i_o}$ -factors at any time  $t \in \mathcal{T}$  are formally identical. Thus we minimize over one factor of the closed-loop model, as the other optimizations are the same, i.e  $c^{i_o} \equiv m^{i_o}(s|a, s')\pi^{i_o}(s|a, s')$ ,  $s, s' \in \mathcal{S}, a \in \mathcal{A}$ . Thus the  $c^{i_o}$  can be formally decomposed as  $c^{i_o} \equiv m^{i_o}\pi^{i_o}$ , similarly to (1.12), then the optimal ideal pd fulfills following relation

$$c^{i_o} \in \arg \min_{\pi^i \in \Pi} \left[ \max_{m^i \in \mathcal{M}^i} \sum_{a \in \mathcal{A}} \pi^i(a|s) \exp[-\mathbf{d}(a, s)] \right], s \in \mathcal{S}, \quad (1.22)$$

that exploits the form of  $\min(\mathbf{D})$  at the considered time  $t$ , see (1.19). Using the (1.22) we can formulate the way how to obtain the  $m^{i_o}$  in following theorem.

**Theorem 4.** Let  $\pi^i \in \Pi^i$  define non-empty cross-section  $\mathcal{M}^i$ . Let  $m^i(s_t|a_t, s_{t-1}) \in \mathcal{M}^i$  exists giving  $\mathbf{d}(a_t, s_{t-1}) < \infty, \forall a_t \in \mathcal{A}, \forall s_{t-1} \in \mathcal{S}$ . Then the optimal  $m^{i_o}$ -factor minimizes  $\mathbf{d}(a_t, s_{t-1})$ , i.e.

$$m^{i_o}(s_t|a_t, s_{t-1}) \in \arg \min_{m^i \in \mathcal{M}^i} \sum_{a \in \mathcal{A}} \pi^i(a_t|s_{t-1}) \exp[-\mathbf{d}(a_t, s_{t-1})] = \arg \min_{m^i \in \mathcal{M}^i} \mathbf{d}(a_t, s_{t-1}), \quad (1.23)$$

$$\forall s_t \in \mathcal{S}, \forall a_t \in \mathcal{A}, \forall s_{t-1} \in \mathcal{S}.$$

*Proof.* For the proof of Theorem 4 see [8]. □

In this part we take a closer look at finding optimal ideal  $\pi^{i_o}$ -factor. From the nature of the FPD-optimal  $\pi^o$  it implies that  $\text{supp}(\pi^o) \subseteq \text{supp}(\pi^i)$ , see (1.18).

We want guarantee that,  $a \in \mathcal{A}$  is a priori excluded.

$$\pi^i \in \Pi^i \equiv \left\{ \pi^i : \text{supp}(\pi^i) = \mathcal{A} \right\} \quad (1.24)$$

The optimal ideal meeting this can be obtained using upcoming theorem.

**Theorem 5.** Let  $\Pi^i$  be given by an opted  $k > 1$  as follows

$$\Pi^i \equiv \left\{ \pi^i : \text{supp}(\pi^i) = \mathcal{A} \text{ and } \|\pi^i\|_k < \infty \right\}, \quad (1.25)$$

with an assumption of  $|\mathcal{A}| < \infty$ .

Let each  $\pi^i$  from (1.25) define non-empty cross-section  $m^{i_o}$  from (1.23). Let  $m^i(s_t|a_t, s_{t-1})$  exist such as  $\mathbf{d}(a_t, s_{t-1}) < \infty, \forall a_t \in \mathcal{A}, \forall s_{t-1} \in \mathcal{S}$ . Then the optimal ideal  $\pi^{i_o}$ -factor is

$$\pi^{i_o}(a_t|s_{t-1}) \propto 1_{\mathcal{A}}(a_t) \exp[-\mu \mathbf{d}^o(a_t, s_{t-1})], \quad \mu \equiv \frac{1}{(k-1)}, \quad (1.26)$$

where

$$\mathbf{d}^o(a_t, s_{t-1}) \equiv \sum_{s \in \mathcal{S}} m(s|a_t, s_{t-1}) \ln \left( \frac{m(s|a_t, s_{t-1})}{h(s)m^{i_o}(s|a_t, s_{t-1})} \right) \leq \mathbf{d}(a_t, s_{t-1}). \quad (1.27)$$

The  $\pi^{i_o}$ -factor is contained in (1.25).

*Proof.* To see proof of Theorem 5, refer to [8]. □

**Remark 6.** The value of  $k$  from the assumption in Theorem 5, respectively value of  $\mu$  influences the exploration rate in PE.

Let us now assume more specific preferences of the agent. He wants to obtain  $s \in \mathcal{S}^i$  and prefers use of  $a \in \mathcal{A}^i$ .

The optimal ideal  $\pi^{i_0}$ -factor is uniquely given by the  $m^{i_0}$ , symbolically written as  $\pi^{i_0} \equiv \pi^{i_0}(m^{i_0})$ , and by the value  $\mu > 1$ . This approach ensures us that agent's preferences are fulfilled, when  $m^{i_0} \in \mathcal{M}^i$  respecting the selected preferences.

The optimal ideal rule  $\pi^{i_0}(m^{i_0})$  then can be written in the following form:

$$\pi^{i_0}(m^{i_0}) \in \arg \max_{m^i \in \mathcal{M}^i} \sum_{a \in \mathcal{A}} \rho(a, s_{t-1}) \pi^i(a|s_{t-1}) \equiv \arg \max_{m^i \in \mathcal{M}^i} \sum_{a \in \mathcal{A}} \left[ \sum_{s \in \mathcal{S}} 1_{\mathcal{S}^i}(s) m(s|a, s_{t-1}) + w 1_{\mathcal{A}^i}(a) \right] \pi^i(a|s_{t-1}) \quad (1.28)$$

The parameter  $w$  fulfills  $w \geq 0$  and represent the importance of acting in  $\mathcal{A}^i \subset \mathcal{A}$  relatively to reaching  $\mathcal{S}^i \subset \mathcal{S}$ . The relation (1.28) has a solution if ideal sets  $\mathcal{A}^i, \mathcal{S}^i$  are "reachable", that is if  $\rho(a, s_{t-1}) > 0$  on  $a \in \mathcal{A}, s_{t-1} \in \mathcal{S}$ .

Now we can derive the optimal value of  $d^o$  important for evaluating factor  $\pi^{i_0}, m^{i_0}$  and thus  $\pi^o$ .

**Theorem 6.** Let  $\text{supp}(\pi^i) = \mathcal{A}, \|\pi^i\|_p < \infty, p > 1$ , and  $|\mathcal{A}| < \infty$ . Let the assumption  $\rho(a, s_{t-1}) > 0$  be met. Then, the optimal ideal  $m^{i_0}$  meeting (1.28) provides  $d^o(a, s)$ , giving  $\pi^{i_0} = \pi^{i_0}(m^{i_0})$ , as the new function

$$d^o(a_t, s_{t-1}) = d^o(\bar{a}_t, s_{t-1}) + \ln \left[ \frac{\rho(\bar{a}_t, s_{t-1})}{\rho(a_t, s_{t-1})} \right], \bar{a}_t \in \arg \max_{a_t \in \mathcal{A}} (\rho(a_t, s_{t-1})), a_t \in \mathcal{A}, s_{t-1} \in \mathcal{S}. \quad (1.29)$$

*Proof.* The proof of Theorem 6 can be found in [8] □

The optimal ideal  $m^{i_0}$  can now be derived giving  $\pi^{i_0}$  via (1.28).

**Theorem 7.** Let  $m(s_t|a_{t-1}, s_{t-1}), a_t \in \mathcal{A}, s_{t-1} \in \mathcal{S}$ , be non-uniform on  $s_t \in \mathcal{S}$  and theorem 5. Then the  $m^{i_0}$ -factor meeting (1.28) has the form

$$m^{i_0}(s_t|a_t, s_{t-1}) \propto m(s_t|a_t, s_{t-1}) \exp[-\mathbf{e}(a_t, s_{t-1}) m(s_t|a_t, s_{t-1})], \quad (1.30)$$

defined under the assumption  $|\mathcal{S}| < \infty$ .

The real-valued  $\mathbf{e}(a_t, s_{t-1})$  in (1.30) is the existing solution of the equation  $L(\mathbf{e}(a_t, s_{t-1})) = R(a_t, s_{t-1})$ , where

$$L(\mathbf{e}(a_t, s_{t-1})) \equiv \mathbf{e}(a_t, s_{t-1}) \Lambda(a_t, s_{t-1}) + \ln \left[ \sum_{s \in \mathcal{S}} m(s|a_t, s_{t-1}) \exp[-\mathbf{e}(a_t, s_{t-1}) m(s|a_t, s_{t-1})] \right], \quad (1.31)$$

$$\Lambda(a_t, s_{t-1}) \equiv \sum_{s_t \in \mathcal{S}} m^2(s_t|a_t, s_{t-1}) > 0, \quad (1.32)$$

$$R(a_t, s_{t-1}) \equiv \sum_{s \in \mathcal{S}} m(s|a_t, s_{t-1}) \ln(h(s)) + d^o(\bar{a}_t, s_{t-1}) + \ln \left[ \frac{\rho(\bar{a}_t, s_{t-1})}{\rho(a_t, s_{t-1})} \right] \geq 0. \quad (1.33)$$

**Remark 7.** For uniform pd  $m(s|a_t, s_{t-1})$  on  $s \in \mathcal{S}$  with  $|\mathcal{S}| < \infty$ , the optimal  $m^{i_0}$  has the form

$$m^{i_0} \propto \exp[-\mathbf{e}(a_t, s_{t-1}) \mathbf{o}(s_t|a_t, s_{t-1})] \quad (1.34)$$

for any non-zero  $\mathbf{o}(s|a_t, s_{t-1})$ , which fulfills assumption  $\int_{\mathcal{S}} \mathbf{o}(s|a_t, s_{t-1}) ds = 0$ . The real-valued  $\mathbf{e}(a_t, s_{t-1})$  is the existing solution of following equation

$$L(\mathbf{e}(a_t, s_{t-1})) \equiv \ln \left[ \frac{\sum_{s \in \mathcal{S}} \exp[-\mathbf{e}(a_t, s_{t-1}) \mathbf{o}(s|a_t, s_{t-1})]}{|\mathcal{S}|} \right] = R(a_t, s_{t-1}) \quad (1.35)$$

$$R(a_t, s_{t-1}) \equiv d^o(\bar{a}_t, s_{t-1}) + \sum_{s \in \mathcal{S}} m(s|a_t, s_{t-1}) \ln \left[ \frac{h(s) \rho(\bar{a}_t, s_{t-1})}{\rho(a_t, s_{t-1})} \right]. \quad (1.36)$$

*Proof.* Proofs of Theorem 7 and Remark 7 can be found in [8] in section 3. □

# Chapter 2

## FPD with Stopping

This section exploits the theory presented in Section 1.3 and extends it by expanding the action and state spaces by stopping related values. As a result of this extension, we needed to properly formulate FPD to reflect stopping actions and states.

### 2.1 Design of the Extended Behaviour

First, we present the extension of the action space, and then we focus on the state space.

**Definition 22.** The newly established *stopping action* is denoted as  $\tilde{a}_t$  and it is defined as follows

$$\tilde{a}_t = \begin{cases} 1 & \text{continue in generating the regular action } a_t, \\ 0 & \text{take the final regular action } a_t \text{ and make an immediate change in generating actions.} \end{cases} \quad (2.1)$$

By this immediate change in generating actions, we understand stopping the process or following another decision rule. This mirrors the classic DM with stopping that does not continue after the stopping, see [19], [14].

**Remark 8.** The newly defined stopping action is joined with the regular action, creating new extended action space with stopping. The extended actions with stopping are then defined as

$$(a_t, \tilde{a}_t) \text{ for } t \in \mathcal{T}.$$

From the definition 22, it is clear that the stopping action  $\tilde{a}_t$  can influence all future behaviour up to the finite horizon  $|\mathcal{T}|$ . We have to turn our attention to the state space and its extension to incorporate the stopping states as well.

**Definition 23.** The new *stopping state* is denoted by  $\tilde{s}_t$  and is defined as

$$\tilde{s}_t = \begin{cases} 1 & \text{the DM process continues,} \\ 0 & \text{the DM is stopped.} \end{cases} \quad (2.2)$$

In the (2.2)  $\tilde{s}_t = 1$  stands for continuation of the DM process in a way similar to the regular DM process without stopping. Whereas  $\tilde{s}_t = 0$  represents that the process is stopped, i.e.  $p(s_{t+1}|s_t) = \delta(s_{t+1}, s_t)$ .

**Remark 9.** The newly defined stopping state is joined with the regular state in the same way as in Remark 8, forming together extended states with stopping as

$$(s_t, \tilde{s}_t) \text{ for } t \in \mathcal{T},$$

where  $s_t$  is the original state generated by the closed loop and  $\tilde{s}_t$  reflects, whether the DM process has stopped or not at time  $t$ .



Now when we have established extended actions with stopping and extended states with stopping we can focus on transforming of the behaviour from Definition 16.

**Definition 24.** The *behaviour of closed loop with stopping* up to the finite horizon  $|\mathcal{T}| \in \mathbb{N}$  is

$$\bar{b} \equiv (s_{|\mathcal{T}|}, \tilde{s}_{|\mathcal{T}|}, a_{|\mathcal{T}|}, \tilde{a}_{|\mathcal{T}|}, s_{|\mathcal{T}|-1}, \dots, a_1, \tilde{a}_1, s_0, \tilde{s}_0) \in \bar{\mathcal{B}}, \quad (2.3)$$

where  $s_t \in \mathcal{S}$ ,  $\tilde{s}_t \in \{0, 1\}$  for  $\forall t \in \{0\} \cup \mathcal{T}$  and  $a_t \in \mathcal{A}$ ,  $\tilde{a}_t \in \{0, 1\}$  for  $\forall t \in \mathcal{T}$  and  $\bar{\mathcal{B}}$  denotes the *set of all behaviours of closed loop with stopping*.

**Remark 10.** The extended form of Definition 17 of the closed-loop model can be rewritten as

$$c^\pi(\bar{b}) = \prod_{t \in \mathcal{T}} m(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) \pi(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}). \quad (2.4)$$

The model in (2.4) can be factorized in the following way

$$m(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = m(s_t | \tilde{s}_t, a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) m(\tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}). \quad (2.5)$$

We propose those individual elements of this factorization to have following forms

$$m(s_t | \tilde{s}_t, a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = [m(s_t | a_t, s_{t-1})]^{\tilde{s}_t} [\delta(s_t, s_{t-1})]^{1-\tilde{s}_t}, \quad (2.6)$$

$$m(\tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = \delta(\tilde{s}_t, \tilde{a}_t). \quad (2.7)$$

**Remark 11.** In equation (2.6), we omitted  $\tilde{s}_{t-1}$  on the right-hand side because it does not have direct impact on the state  $s_t$  if the process continued, but it influences  $\tilde{a}_t$  and  $\tilde{s}_t$ , which then directly influences generating of the next state  $s_t$ . Part  $[\delta(s_t, s_{t-1})]^{1-\tilde{s}_t}$  represents generating of states in our process after we decided that it should be stopped, i.e.  $\tilde{s}_t = 0$ .

**Remark 12.** The formula (2.7) was designed in this way, because we want to have stopping under our full control. If we take the decision to stop  $\tilde{a} = 0$  we want the process to be stopped immediately. This specific formula could be designed in a different way, but it would lead to a different solution.

**Remark 13.** By using ordinary chain rule we can obtain also a different decomposition, in this work we use the one given above (2.5) due to its closeness to the model from Definition 17.

Now we can present our extended decision rule.

**Remark 14.** The *extended decision rule* can be decomposed as

$$\pi(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) = \pi(a_t | \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) \pi(\tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}). \quad (2.8)$$

The first factor  $\pi(a_t | \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})$  is defined as

$$\pi(a_t | \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = [\pi(a_t | s_{t-1})]^{\tilde{a}_t} [\tilde{\pi}(a_t | s_{t-1})]^{1-\tilde{a}_t}. \quad (2.9)$$

It reflects the fact that after deciding to stop  $\tilde{a}_t = 0$ , there can be made some change in generating of regular actions described by an optional  $\tilde{\pi}$ , e.g. regular actions are optimized, generated only randomly or are fixed.

The second factor  $\pi(\tilde{a}_t | s_{t-1}, \tilde{s}_{t-1})$  in (2.8) represents the decision rule determining whether the process should stop or continue. The meaning of  $\tilde{a}_t$  implies the form of this decision rule

$$\pi(\tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) = \begin{cases} 1 & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 0, \\ 0 & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 0, \\ 1 - \mathbf{q}(s_{t-1}) & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 1, \\ \mathbf{q}(s_{t-1}) & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 1. \end{cases} \quad (2.10)$$

This decision rule expresses, that if the process is already stopped  $\tilde{s}_{t-1} = 0$  it cannot continue. Then, there is a part which contains the so-called *prolonging probability* represented by  $\mathbf{q}(s_{t-1}) \in [0, 1]$ , which is the optional parameter of this decision rule. It is probability to continue if the process is in the state  $s_{t-1}$ . The values of  $\mathbf{q}(s_{t-1})$  should be designed in the way that reflects the willingness to stop or continue at certain state.

## 2.2 Design of the Extended Ideal Probability

Using of FPD approach requires an ideal probability of behaviour of the closed loop that represents our preferences.

**Remark 15.** The extended ideal probability of behaviours of the closed loop from Definition 18 can be written as

$$c^i(\bar{b}) = \prod_{t \in \mathcal{T}} m^i(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) \pi^i(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}). \quad (2.11)$$

The *ideal model with the stopping* from (2.11) can be decomposed in the following way

$$m^i(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = m^i(s_t | \tilde{s}_t, a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) m^i(\tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}). \quad (2.12)$$

The first factor in (2.12) is designed in a following way

$$m^i(s_t | \tilde{s}_t, a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = [m^i(s_t | a_t, s_{t-1})]^{\tilde{s}_t} [m(s_t | a_t, s_{t-1})]^{1-\tilde{s}_t}. \quad (2.13)$$

This form reflects that the *original<sup>1</sup> ideal model* is used while the process is continuing,  $\tilde{s}_t = 1$ . If the process is stopped,  $\tilde{s}_t = 0$ , we switch the ideal model to a non-ideal model because the DM process has been stopped and our preferences do not matter.

The second factor in (2.12) is in the simple form

$$m^i(\tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = \delta(\tilde{s}_t, \tilde{a}_t). \quad (2.14)$$

The reasoning leading to this form is the same as in Remark 12.

**Remark 16.** The *ideal decision rule* can be factorized in similar way as *decision rule* in Remark 14 as follows

$$\pi^i(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) = \pi^i(a_t | \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) \pi^i(\tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}). \quad (2.15)$$

The chosen form of the first factor is

$$\pi^i(a_t | \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = [\pi^i(a_t | s_{t-1})]^{\tilde{s}_t} [\tilde{\pi}(a_t | s_{t-1})]^{1-\tilde{s}_t}, \quad (2.16)$$

where  $\pi^i(a_t | s_{t-1})$  is the *original ideal decision rule*,  $\tilde{\pi}(a_t | s_{t-1})$  is a unspecified<sup>2</sup> *decision rule*. The second factor in (2.15) reads

$$\pi^i(\tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) = \begin{cases} 1 & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 0, \\ 0 & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 0, \\ 1 - q^i(s_{t-1}) & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 1, \\ q^i(s_{t-1}) & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 1, \end{cases} \quad (2.17)$$

where the ideal probability of continuation  $q^i(s_{t-1})$  reflects the price of prolonged action.

**Remark 17.** The values 1 and 0 in (2.17) are correct because they satisfy one of the main restrictions given on stopping. This restriction is that if we change the process once  $\tilde{s}_t = 0$ , it cannot be undone, so the upcoming  $\tilde{a}_\tau = 0 \forall \tau > t$ .

<sup>1</sup>Original means in the DM task without stopping.

<sup>2</sup>For instance,  $\tilde{\pi}(a_t | s_{t-1})$  can be uniform probability.

## 2.3 Evaluation of the Optimal Policy using FPD with Stopping

Finally, after the previous extension of the joint pd of behaviours and related pds, we can solve FPD with stopping. Thus, we can present our extension of Theorem 3.

**Theorem 8.** (Solution of FPD with Stopping) The optimal policy in the closed-loop process with stopping can be evaluated by using following form

$$\pi^o(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) = \pi^i(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) \frac{\exp[-\mathbf{d}(a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})]}{h(s_{t-1}, \tilde{s}_{t-1})}, \quad (2.18)$$

where

$$h(s_{t-1}, \tilde{s}_{t-1}) = \sum_{\substack{a_t \in \mathcal{A} \\ \tilde{a}_t \in \{0,1\}}} \pi^i(a_t, \tilde{a}_t | s_{t-1}, \tilde{s}_{t-1}) \exp[-\mathbf{d}(a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})] \quad (2.19)$$

and

$$\mathbf{d}(a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = \sum_{\substack{s_t \in \mathcal{S} \\ \tilde{s}_t \in \{0,1\}}} m(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) \ln \left( \frac{m(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})}{m^i(s_t, \tilde{s}_t | a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) h(s_t, \tilde{s}_t)} \right). \quad (2.20)$$

The above given formulas are only the extended forms of (1.19) and (1.20).

We have to add an necessary assumption, which enables us to find the solution using backward recursion. This assumption necessary for initialization of the backward recursion is

$$h(s_{|\mathcal{T}|}, \tilde{s}_{|\mathcal{T}|}) = 1, \forall s_{|\mathcal{T}|} \in \mathcal{S}, \tilde{s}_{|\mathcal{T}|} \in \{0, 1\}. \quad (2.21)$$

*Proof.* The proof of Theorem 8 is identical with the proof of Theorem 3, but instead of action  $a_t$  and states  $s_t$  in Theorem 3 we deal with  $(a_t, \tilde{a}_t)$  and  $(s_t, \tilde{s}_t)$ .  $\square$

**Remark 18.** The assumption (2.21) is extended version of  $h(s_{|\mathcal{T}|}) = 1, \forall s_{|\mathcal{T}|} \in \mathcal{S}$  in Theorem 3.

Now we can derive functions  $\mathbf{d}(\bullet)$  and  $h(\bullet)$  which are required for calculation of the optimal rule with stopping.

If we look closely to assumption (2.21) and combine it with the fact, that already stopped process does not develop further, i.e.  $\tilde{s}_t = 0 \Rightarrow s_t = s_{t-1}$ . We get the following identities

$$h(s_{t-1}, \tilde{s}_{t-1} = 0) = h(s_t, \tilde{s}_t = 0) = \dots = h(s_{|\mathcal{T}|}, \tilde{s}_{|\mathcal{T}|} = 0) = 1, \text{ for } \forall s_{t-1} \in \mathcal{S}. \quad (2.22)$$

Now we move to calculation of the function  $\mathbf{d}(\bullet)$ . We start with the evaluation of  $\mathbf{d}(a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0)$ . For this evaluation we will be using following formulas (2.5), (2.6) and (2.7) for extended model, (2.12), (2.13) and (2.14) for extended ideal model and finally the condition (2.22).

$$\begin{aligned} \mathbf{d}(a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0) &= \sum_{\substack{s_t \in \mathcal{S} \\ \tilde{s}_t \in \{0,1\}}} m(s_t, \tilde{s}_t | a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0) \\ &\quad \ln \left( \frac{m(s_t, \tilde{s}_t | a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0)}{m^i(s_t, \tilde{s}_t | a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0) h(s_t, \tilde{s}_t)} \right) = \end{aligned} \quad (2.23)$$

$$= \sum_{s_t \in \mathcal{S}} \delta(s_t, s_{t-1}) \delta(\tilde{s}_t = 0, \tilde{a} = 0) \ln \left( \frac{\delta(s_t, s_{t-1}) \delta(\tilde{s}_t = 0, \tilde{a} = 0)}{\delta(s_t, s_{t-1}) \delta(\tilde{s}_t = 0, \tilde{a} = 0) h(s_t, \tilde{s}_t = 0)} \right) = 0. \quad (2.24)$$

By the use of (2.7) we reduced the sum to the case when  $\tilde{s}_t = 0$ , because  $m(s_t, \tilde{s}_t = 1|a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0) = 0, \forall s_t \in \mathcal{S}$ . Then, we inserted correct forms of extended model and ideal model and we obtain last equality (2.24).

Finally with use of the condition (2.22) we can see, that

$$\mathbf{d}(a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 0) = 0, \forall a_t \in \mathcal{A}, s_t \in \mathcal{S}. \quad (2.25)$$

Evaluation of  $\mathbf{d}(a_t, \tilde{a}_t = 1, s_{t-1}, \tilde{s}_{t-1} = 0)$  is not necessary for the evaluation of the optimal policy, because in evaluation of the optimal policy in (2.18) and in (2.19) is multiplied by  $\pi^i(a_t, \tilde{a}_t = 1|s_{t-1}, \tilde{s}_{t-1} = 0)$  which is as we can see in (2.16) equal to 0.

Next we evaluate  $\mathbf{d}(a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)$ .

$$\begin{aligned} \mathbf{d}(a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1) &= \sum_{\substack{s_t \in \mathcal{S} \\ \tilde{s}_t \in \{0,1\}}} m(s_t, \tilde{s}_t|a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1) \\ &\quad \ln \left( \frac{m(s_t, \tilde{s}_t|a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)}{m^i(s_t, \tilde{s}_t|a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)h(s_t, \tilde{s}_t)} \right) = \end{aligned} \quad (2.26)$$

$$\begin{aligned} &= \sum_{\substack{s_t \in \mathcal{S} \\ \tilde{s}_t \in \{0,1\}}} m(s_t|\tilde{s}_t, a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)\delta(\tilde{s}_t, \tilde{a}_t = 0) \\ &\quad \ln \left( \frac{m(s_t|\tilde{s}_t, a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)\delta(\tilde{s}_t, \tilde{a}_t = 0)}{m^i(s_t|\tilde{s}_t, a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)\delta(\tilde{s}_t, \tilde{a}_t = 0)h(s_t, \tilde{s}_t)} \right) = \end{aligned} \quad (2.27)$$

$$\begin{aligned} &= \sum_{s_t \in \mathcal{S}} \delta(s_t, s_{t-1})\delta(\tilde{s}_t = 0, \tilde{a}_t = 0) \ln \left( \frac{\delta(s_t, s_{t-1})\delta(\tilde{s}_t = 0, \tilde{a}_t = 0)}{\delta(s_t, s_{t-1})\delta(\tilde{s}_t = 0, \tilde{a}_t = 0)h(s_t, \tilde{s}_t = 0)} \right) = \\ &= \sum_{s_t \in \mathcal{S}} \delta(s_t, s_{t-1}) \ln \left( \frac{\delta(s_t, s_{t-1})}{\delta(s_t, s_{t-1})h(s_t, \tilde{s}_t = 0)} \right) = 0 \end{aligned} \quad (2.28)$$

Final form which we need to evaluate is  $\mathbf{d}(a_t, \tilde{a}_t = 1, s_{t-1}, \tilde{s}_{t-1} = 1)$  and its final form is as follows

$$\mathbf{d}(a_t, \tilde{a}_t = 1, s_{t-1}, \tilde{s}_{t-1} = 1) = \sum_{s_t \in \mathcal{S}} m(s_t|a_t, s_{t-1}) \ln \left( \frac{m(s_t|a_t, s_{t-1})}{m^i(s_t|a_t, s_{t-1})h(s_t, \tilde{s}_t = 1)} \right). \quad (2.29)$$

Finally we focus on the examination of  $h(s_{t-1}, \tilde{s}_{t-1} = 1)$ .

$$h(s_{t-1}, \tilde{s}_{t-1} = 1) = \sum_{a_t \in \mathcal{A}} \pi^i(a_t, \tilde{a}_t = 0|s_{t-1}, \tilde{s}_{t-1} = 1) \exp[-\mathbf{d}(a_t, \tilde{a}_t = 0, s_{t-1}, \tilde{s}_{t-1} = 1)] + \quad (2.30)$$

$$+ \sum_{a_t \in \mathcal{A}} \pi^i(a_t, \tilde{a}_t = 1|s_{t-1}, \tilde{s}_{t-1} = 1) \exp[-\mathbf{d}(a_t, \tilde{a}_t = 1, s_{t-1}, \tilde{s}_{t-1} = 1)] = \quad (2.31)$$

$$= \sum_{a_t \in \mathcal{A}} \pi^i(a_t|s_{t-1}) (1 - \mathbf{q}^i(s_{t-1})) + \sum_{a_t \in \mathcal{A}} \pi^i(a_t|s_{t-1}) \mathbf{q}^i(s_{t-1}) \exp[-\mathbf{d}(a_t, \tilde{a}_t = 1, s_{t-1}, \tilde{s}_{t-1} = 1)] \quad (2.32)$$

Now when we have all necessary forms of  $\mathbf{d}(\bullet)$ ,  $h(\bullet)$  and ideal decision rule with stopping from remark 16, we can present final for of our optimal policy:

$$\pi^o(a_t, \tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}) \propto \begin{cases} \tilde{\pi}(a_t|s_{t-1}) & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 0, \\ 0 & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 0, \\ \pi^i(a_t|s_{t-1})[1 - \mathbf{q}^i(s_{t-1})] \frac{1}{h(s_{t-1}, \tilde{s}_{t-1}=1)} & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 1, \\ \pi^i(a_t|s_{t-1}) \mathbf{q}^i(s_{t-1}) \frac{\exp[-\mathbf{d}(a_t, \tilde{a}_t=1, s_{t-1}, \tilde{s}_{t-1}=1)]}{h(s_{t-1}, \tilde{s}_{t-1}=1)} & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 1. \end{cases} \quad (2.33)$$

**Remark 19.** In this remark we discuss how to compute the optimal rule using backward recursion used on a selected sufficiently large time horizon  $|\mathcal{T}|$ .

For this we will use previously derived relations. It can be seen from Theorem 8, that relation (2.19) is inserted into relation (2.20). From that it is obvious that assumption needed for the initialization is the one, which initializes value  $h(s_{|\mathcal{T}|}, \tilde{s}_{|\mathcal{T}|})$ .

As we can see above, if we prepare all the needed components and add the initialization step, we can begin our backward recursion to compute the values of our functions  $h(\bullet)$  and  $d(\bullet)$ , from which we derive the optimal policy  $\pi^o$ . That initialization step is the one already mentioned in Theorem 8 and has the following form

$$h(s_{|\mathcal{T}|}, \tilde{s}_{|\mathcal{T}|}) = 1, \forall s_{|\mathcal{T}|} \in \mathcal{S}, \tilde{s}_{|\mathcal{T}|} \in \{0, 1\}.$$

## Chapter 3

# Preference Elicitation in FPD with Stopping

In this chapter is presented the incorporation of the PE from Section 1.4 into formalism derived in Chapter 2. This would lead to easier work for DM designer, by removing necessity for designing ideal probabilities  $m^i$  in (2.20) and related formulas.

### 3.1 Algorithm for PE

For the illustration purpose we show here the overview of the algorithm we use for evaluation of  $m^{i_0}$  and  $\pi^{i_0}$ .

---

**Algorithm 1** PE in FPD with stopping algorithm

---

1: **Inputs:**

- finite state space  $\mathcal{S}$  and finite action space  $\mathcal{A}$ , ideal state space  $\mathcal{S}^i$  and ideal action space  $\mathcal{A}^i$
- System model  $m(s|a, \bar{s}), \forall s, \bar{s} \in \mathcal{S}, a \in \mathcal{A}$
- Selected weight  $w \geq 0$ , representing preferences between  $\mathcal{S}^i$  and  $\mathcal{A}^i$
- Design horizon  $|\mathcal{T}|$ , exploration parameter  $\mu > 1$  and an initial step for backward recursion  $h(s_{|\mathcal{T}|}) = 1, \forall s_{|\mathcal{T}|} \in \mathcal{S}$

2: **Evaluation of all auxiliary functions**3: **for all**  $\forall s_{t-1} \in \mathcal{S}$  **do**4:   **for all**  $\forall a_t \in \mathcal{A}$  **do**

5:      $\rho(a_t, s_{t-1}) = \sum_{s \in \mathcal{S}} 1_{\mathcal{S}^i}(s) m(s|a_t, s_{t-1}) + w 1_{\mathcal{A}^i}(a_t)$

6:      $\Lambda(a_t, s_{t-1}) = \sum_{s \in \mathcal{S}} m^2(s|a_t, s_{t-1})$

7:   **end for**

8:    $(\bar{a}_t) \in \arg \max_{a_t} \rho(a_t, s_{t-1})$

9:    $\bar{\rho} = \rho(\bar{a}_t, s_{t-1})$

10: **end for**

11:  $t \leftarrow |\mathcal{T}|$

12: **while**  $t > 0$  **do**13:   **Evaluation of  $d^o$  function**14:   **for all**  $\forall s_{t-1} \in \mathcal{S}$  **do**15:     **for all**  $\forall a_t \in \mathcal{A}$  **do**

16:        $d^o(\bar{a}_t, s_{t-1}) = \max \left[ 0, \max_{a_t \in \mathcal{A}} \sum_{s \in \mathcal{S}} m(s|a_t, s_{t-1}) \ln \left( \frac{\rho(a_t, s_{t-1})}{h(s)\bar{\rho}} \right) \right]$

17:        $d^o(a_t, s_{t-1}) = d^o(\bar{a}_t, s_{t-1}) + \ln \left( \frac{\bar{\rho}}{\rho(a_t, s_{t-1})} \right)$

18:     **if**  $m(s_t|a_t, s_{t-1})$  is non-uniform **then**

19:        $R(a_t, s_{t-1}) = \sum_{s \in \mathcal{S}} m(s|a_t, s_{t-1}) \ln(h(s)) + d^o(\bar{a}_t, s_{t-1}) + \ln \left[ \frac{\bar{\rho}}{\rho(a_t, s_{t-1})} \right]$

20:        $L(e(a_t, s_{t-1})) = e(a_t, s_{t-1}) \Lambda(a_t, s_{t-1}) + \ln \left[ \sum_{s \in \mathcal{S}} m(s_t|a_t, s_{t-1}) \exp[-e(a_t, s_{t-1}) m(s_t|a_t, s_{t-1})] \right]$

21:       Find an existing real-valued solution  $e(a_t, s_{t-1})$  in

22:        $R(a_t, s_{t-1}) = L(e(a_t, s_{t-1}))$

23:        $m^{i_o}(s_t|a_t, s_{t-1}) \propto m(s_t|a_t, s_{t-1}) \exp[-e(a_t, s_{t-1}) m(s_t|a_t, s_{t-1})]$

24:     **end if**

25:      $\pi^{i_o}(a_t|s_{t-1}) = \exp[-\mu d^o(a_t, s_{t-1})]$

26:   **end for**27:   Normalize  $\pi^{i_o}(a_t|s_{t-1})$ 

28:    $N(s_{t-1}) = \sum_{a \in \mathcal{A}} \pi^{i_o}(a_t|s_{t-1}) \exp[-d^o(a_t, s_{t-1})]$

29:    $\pi^o(a_t|s_{t-1}) = \frac{\exp[-(\mu + 1)d^o(a_t, s_{t-1})]}{N(s_{t-1})}, \forall a_t \in \mathcal{A}$

30: **end for**

31:  $h(s_t) = N(s_{t-1}), \forall s_t \in \mathcal{S}$

32:  $t \leftarrow t - 1$

33: **end while**34: **Outputs:**  $m^{i_o}, \pi^{i_o}, \pi^o$  factors

---

### 3.2 Incorporation of Preference Elicitation into FPD with Stopping

In this section we propose the way of exploiting the PE for solving FPD with stopping. The use of PE relieve necessity of designing  $m^i(s_t, \tilde{s}_t|\cdot)$  and  $\pi^i(a_t|s_{t-1})$ . These probabilities are evaluated using

selected preferred states  $\mathcal{S}^i$ , preferred actions  $\mathcal{A}^i$ , weight  $w$  and exploration parameter  $\mu$  as presented in chapter 1.4.

Using presented PE we can evaluate  $m^{i_o}(s_t|a_t, s_{t-1})$  and  $\pi^{i_o}(a_t|s_{t-1})$ . Those can be used for obtaining ideal model with stopping  $m^i(s_t, \tilde{s}_t|\cdot)$  and ideal decision rule with stopping  $\pi^i(a_t, \tilde{a}_t|\cdot)$  in the following way

$$m^{i_o}(s_t, \tilde{s}_t|a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = [m^{i_o}(s_t|a_t, s_{t-1})]^{\tilde{s}_t} [m(s_t|a_t, s_{t-1})]^{1-\tilde{s}_t} \delta(\tilde{s}_t, \tilde{a}_t) \quad (3.1)$$

$$m(s_t, \tilde{s}_t|a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = m(s_t|a_t, s_{t-1})\delta(\tilde{s}_t, \tilde{a}_t) \quad (3.2)$$

$$\pi^i(a_t, \tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}) = \pi^{i_o}(a_t|s_{t-1})\pi^i(\tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}), \quad (3.3)$$

where  $\pi^i(\tilde{a}_t|s_{t-1}, \tilde{s}_{t-1})$  we designed in already presented form in section 2.2 as follows, but with slight modification

$$\pi^i(\tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}) = \begin{cases} 1 & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 0, \\ 0 & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 0, \\ 1 - q^i & \text{for } \tilde{a}_t = 0 \text{ if } \tilde{s}_{t-1} = 1, \\ q^i & \text{for } \tilde{a}_t = 1 \text{ if } \tilde{s}_{t-1} = 1. \end{cases} \quad (3.4)$$

**Remark 20.** We have made a modification of  $\pi^i(\tilde{a}_t|s_{t-1}, \tilde{s}_{t-1})$  from (2.17) and implemented it in the following way  $q^i(s) = q^i, \forall s \in \mathcal{S}$ , as we do not want to influence our stopping decision of dynamic programming based on previous state.

These formulas can be inserted into Theorem 8, giving us the modified version of the solution of FPD with stopping.

$$\pi^o(a_t, \tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}) = \pi^{i_o}(a_t|s_{t-1})\pi^i(\tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}) \frac{\exp[-\mathbf{d}^{i_o}(a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})]}{h(s_{t-1}, \tilde{s}_{t-1})}, \quad (3.5)$$

where

$$h(s_{t-1}, \tilde{s}_{t-1}) = \sum_{\substack{a_t \in \mathcal{A} \\ \tilde{a}_t \in \{0,1\}}} \pi^{i_o}(a_t|s_{t-1})\pi^i(\tilde{a}_t|s_{t-1}, \tilde{s}_{t-1}) \exp[-\mathbf{d}^{i_o}(a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})] \quad (3.6)$$

and

$$\mathbf{d}^{i_o}(a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) = \sum_{\substack{s_t \in \mathcal{S} \\ \tilde{s}_t \in \{0,1\}}} m(s_t, \tilde{s}_t|a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1}) \ln \left( \frac{m(s_t, \tilde{s}_t|a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})}{m^{i_o}(s_t, \tilde{s}_t|a_t, \tilde{a}_t, s_{t-1}, \tilde{s}_{t-1})h(s_t, \tilde{s}_t)} \right). \quad (3.7)$$

It can be easily derived that formulas from the above leads to the equivalent formula (2.33) for  $\pi^o(a_t, \tilde{a}_t|s_{t-1}, \tilde{s}_{t-1})$ .

### 3.3 Bayesian Parameter Estimation

The FPD and PE inherently rely on knowledge of the model of system for evaluation of the optimal policy. However, in cases where our knowledge of the system's model is insufficient, we must estimate it. One of the methods suitable for the estimation of the system model is Bayesian parameter estimation, which we use in following simulation section. This method is used to illustrate practical use of our proposed solutions from Section 3.2 on real-life example like controlling the system while estimating it.



In this work we use standard version of this estimation which is sufficient for estimation of unknown but time-invariant values of transition probabilities of the system  $\Theta$ . Obtained parametric model  $m(s_t|a_t, s_{t-1}, \Theta)$  is from exponential family, refer to [1]. Counting the observed sequences of  $s_{t-1} \in \mathcal{S}, a_t \in \mathcal{A}$  and  $s_t \in \mathcal{S}$  forms the sufficient statistics for learning the unknown transition probability  $\Theta_{s_t|a_t, s_{t-1}} \equiv m(s_t|a_t, s_{t-1}, \Theta)$  for  $s_t, s_{t-1} \in \mathcal{S}, a_t \in \mathcal{A}$ . For further details about Bayesian approach to estimation of the unknown system, see [13].

The exploratory aspect of FPD enables us to utilize a certainty-equivalent policy, employing point estimates of transition probabilities rather than the use of probabilities themselves.

### Diagram of PE in FPD with Stopping

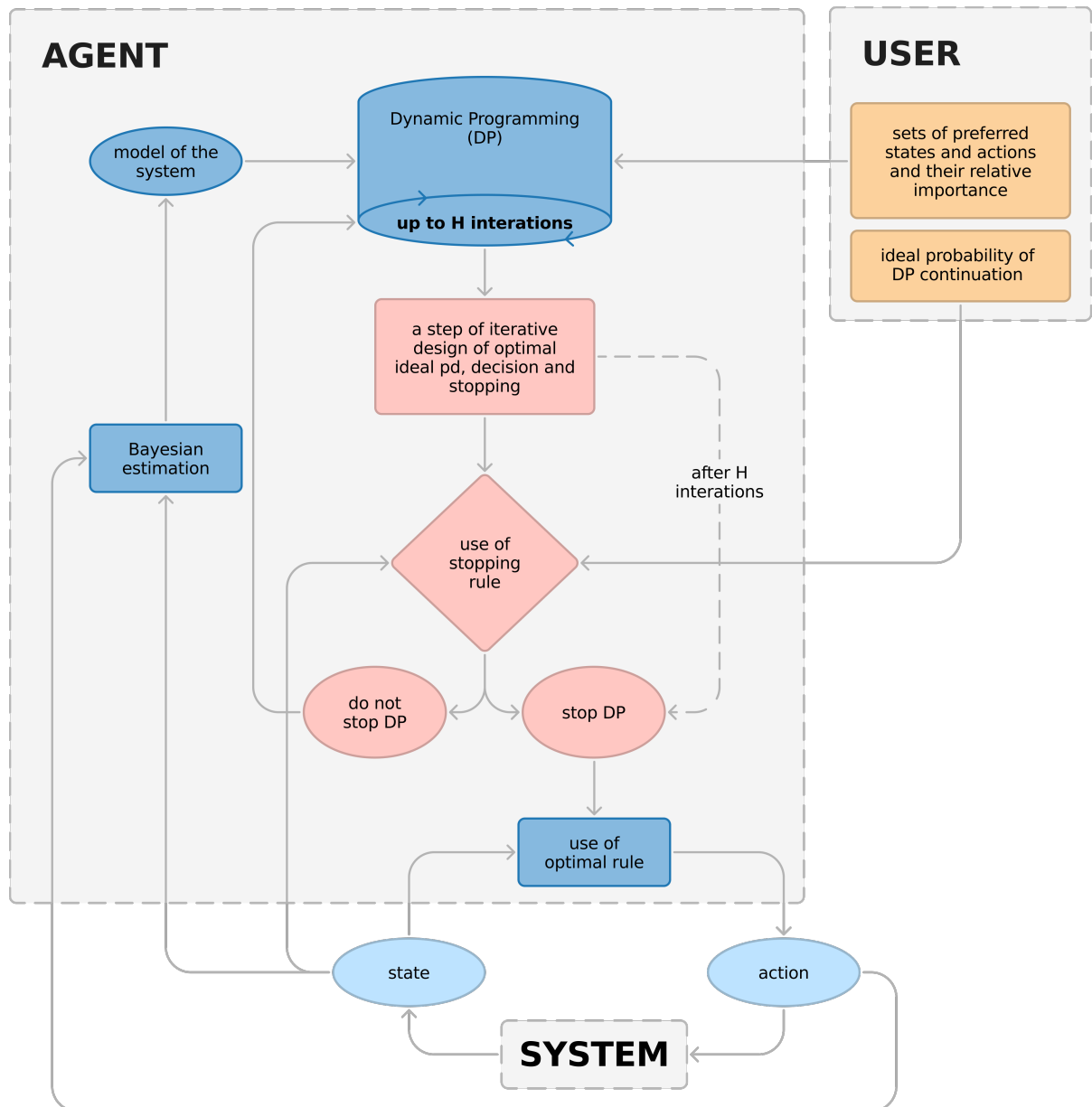


Figure 3.1: Diagram of our proposed solution used in Simulations. User provides external settings to the Agent. With letter H we understand value of horizon. With red color we distinguished parts that was modified or added to the standard solution of FPD.

# Chapter 4

## Simulations

This chapter is dedicated to the presentation of illustrative examples of FPD with stopping using PE as presented in Chapters 2 and 3. Through a series of designed experiments, we aim to demonstrate the performance of our proposed method in different settings. The subsequent sections will provide a detailed account of the experimental setup, simulated unknown system, and a thorough analysis of simulated experiments. The objective is to offer a comprehensive evaluation of the method's effectiveness, considering its strengths, limitations, applicability, and its direct comparison with standard FPD using PE with fixed horizon. By systematically comparing its performance metrics, we seek to provide an objective and transparent assessment of our proposed solution. What we understand as performance metrics is presented in the following Section 4.2.

### 4.1 Design of the Simulation

In this section, we outline the methodology used for generating the unknown transition probabilities of the system that we are aiming to influence in the way to obtain preferred states and using preferred actions. For the purposes of gaining the unknown transition probability of the system, we generated 10000 observations  $y_t$  using following normal linear model

$$y_t = 0.99y_{t-1} + 0.05a_t - 0.125 + 0.05e_t,$$

where  $e_t$  represents white noise with zero mean and unit variance,  $y_{t-1}$  represents previous data point and  $a_t$  represents selected action. The actions  $a_t$  were generated using a uniform distribution over the set  $\{0, 1, 2, 3, 4\}$ . The generation of the data was initialized to  $y_0 = 5.5$ . All generated data points  $y_t$  were then grouped into 11 intervals of exactly the same length with empty intersections. Using this approach we have discretized generated continuous data into 11 possible states.

After this discretization, we added a 1 to 3-dimensional occurrence table of shape  $11 \times 5 \times 11$  for each observation of  $s_t$ , which was influenced by action  $a_t$  and the previous state  $s_{t-1}$ . For practical and numerical purposes, we did not initialize this 3-dimensional occurrence table to 0 values, but we initialized it to values of  $10^{-5}$  to avoid any numerical problems, such as division by zero etc. Finally, the unknown transition probabilities were estimated using the occurrence table, which was normalized to form probabilities, cf. 3.3.

All experiments were run on the simulated system created in this way, and all experiments were initialized to the same state and action.

In the experiments, we used the Monte Carlo method to attain a more comprehensive and complex overview of overall performance of our proposed method. The Monte Carlo method is a powerful tool to simulate diverse scenarios through random sampling. For more in-depth information about Monte Carlo method and its use in various fields of study we refer to [18].

## 4.2 Performance Metrics

In this section, we discuss the key performance metrics used to rigorously evaluate the effectiveness of our proposed method. An understanding of these metrics is important in assessing performance of the method across various dimensions. Each metric serves as a quantitative measure and provides insight into specific aspects of the method's functionality.

The first performance metric we are going to use is the amount of obtained preferred states. This would be closely tied to comparing with the amount of used preferred actions. This metric is easily comparable between our proposed method and regular FPD using PE with same settings. We also compare the counts of realized preferred states and used actions, visualized using histograms.

The second performance metric is a computational time. The time needed for the evaluation of a single run of the experiment. In all our presented examples we have made 50 Monte Carlo repetitions of single runs for each method, both with and without stopping, under the same initial settings. This metric illustrates the time-effectiveness of our dynamic method with stopping compared with the standard one without stopping. This allow us to evaluate the quality of our solution with respect to computational and time costs. We tried to maintain the same computational conditions for every single experiment. The experiments were conducted on a server PC built to maintain stable long-term computing. No other active processes were running on the PC during the computation of the experiments.

The last observed performance metric was the time at which the dynamic programming stopped. The maximum value of this variable can be equal to the value of the design horizon  $|\mathcal{T}|$  and the minimum value is equal to 1.

## 4.3 Input Settings

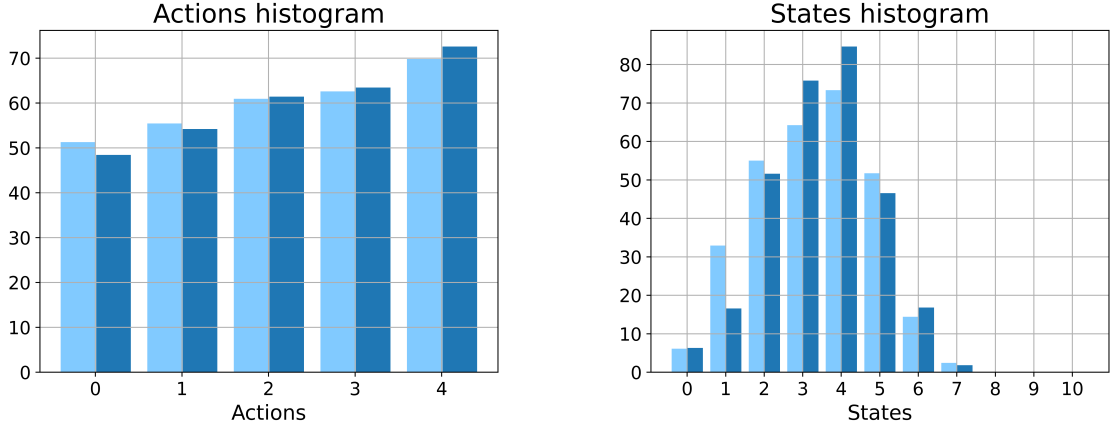
Here we would describe all input variables, which can be selected at the beginning of the experiment. The first variable which can be selected is the number of MC runs, which represents the number of times the single experiment is repeated. The second one is the value of horizon  $|\mathcal{T}|$ , which influences length of the evaluation of the optimal decision rule in PE and in FPD with stopping as well. Then follows settings of variables mentioned in PE as  $\mathcal{S}^i, \mathcal{A}^i, w, \mu$ , see section 1.4. The next variable, which needs to be filled is the length of the single experiment, in other words how many cycles of observing state and generating action, will be done. The last variable is value of  $q^i$ , which represents probability of continuation of the dynamic programming, cf. section 3.2.

The length of the simulation was for all presented experiments equal to 300. The number of Monte Carlo repetitions was for all experiments set to 50. Choices of other input settings are described at respective experiments.

## 4.4 First Experiment

In this experiment we were focusing on a scenario with given preferences only on the state space, with no preferences put on the selection of actions. Our preferred states were  $\mathcal{S}^i = \{4, 5, 6\}$  and  $\mathcal{A}^i = \{0, 1, 2, 3, 4\} = \mathcal{A}$ . Other related parameters were selected as follows  $\mu = 1, w = 0.005$  and  $|\mathcal{T}| = 50$ . Probability of continuation in ideal stopping probability was set to  $q^i = 0.85$ .

First, we will examine histograms of states and actions in our experiment.



(a) Histogram of actions in the experiment with no preferred actions selected. (b) Histogram of states in the experiment with preferred states  $\mathcal{S}^i = \{4, 5, 6\}$ .

Figure 4.1: Experiment with preferred states  $\mathcal{S}^i = \{4, 5, 6\}$  and no preferred actions  $\mathcal{A}^i = \mathcal{A}$ . The input settings were  $\mu = 1$ ,  $w = 0.005$ ,  $q^i = 0.85$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

In Figure 4.1a it can be seen that none action was dominantly selected over the other actions. However in the case with fixed horizon actions 3 and 4 were slightly more used than other actions. That indicates that these actions lead to generating preferred states 4, 5, 6. In the case with dynamic horizon the actions are distributed relatively similarly. But because we did not have any specific action selected as preferred, we will focus on counts of observed preferred states  $\mathcal{S}^i = \{4, 5, 6\}$ . For the fixed horizon preferred states were observed 149 time in total. That is approximately 49.7 % of all states. When we compare it with our proposed method with dynamic horizon we observed preferred states 140 times. That means that in this case 46.7% of all states were preferred states. That is drop in performance around 6.0% in the number of realized preferred states. It is the cost for a significant spearing of time. It is seen on the change in time performance.

Type of Horizon	Median	Mean	Std. Deviation	Max.	Min.
Fixed	44.02	43.94	1.68	47.31	42.11
Dynamic	20.54	20.46	1.08	22.42	17.99

Table 4.1: Statistics of computational time (in seconds) for a single run of the experiment with the input settings  $\mathcal{S}^i = \{4, 5, 6\}$ ,  $\mathcal{A}^i = \{0, 1, 2, 3, 4\}$ ,  $\mu = 1$ ,  $w = 0.005$ ,  $q^i = 0.85$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

From the values in Table 4.1 we can see that our dynamic method did very well in comparison with the method with fixed horizon. Whether we compare mean or median values we can see drop in computational time around 53%, which is really a significant improvement. When we compare other statistics like standard deviation, maximum or minimum, it can be seen that even in these statistics our dynamic method is better.

It is up to every designer or decision maker, what is more important for him, whether quality of final results or speed of computational costs. The ideal probability of continuation  $q^i$  controls this balance.

Finally we display the histogram of stopping time in the first experiment.

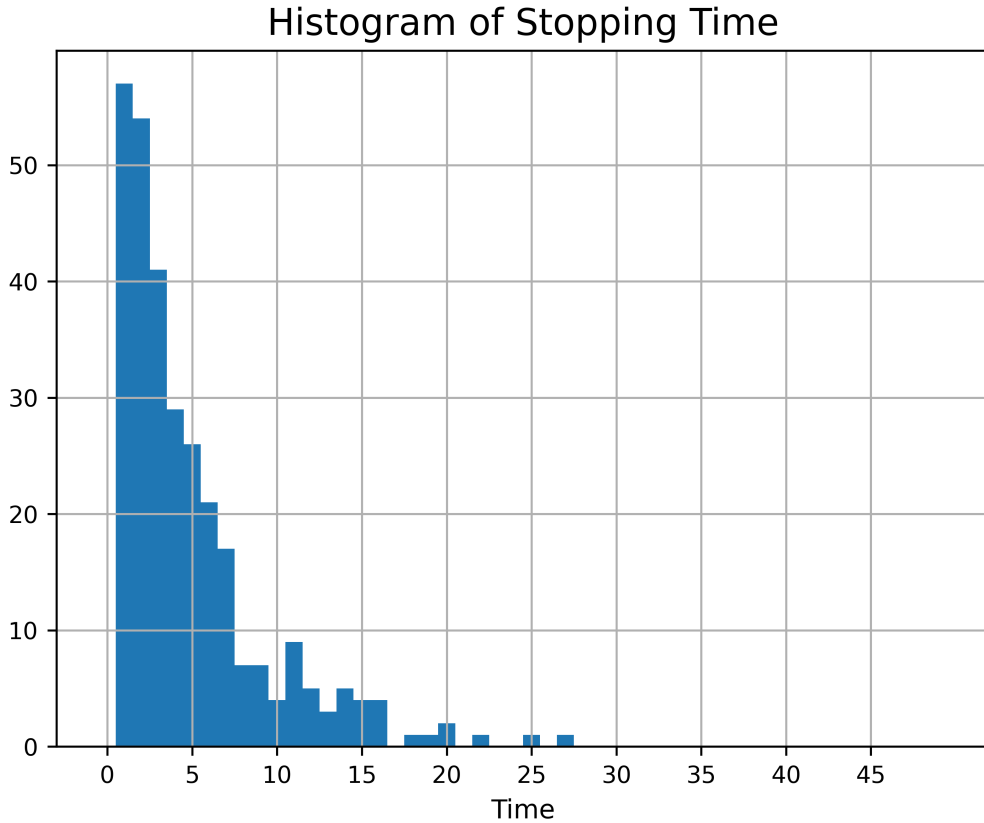


Figure 4.2: Histogram of stopping time of the experiment using input settings  $\mathcal{S}^i = \{4, 5, 6\}$ ,  $\mathcal{A}^i = \{0, 1, 2, 3, 4\}$ ,  $\mu = 1$ ,  $w = 0.005$ ,  $q^i = 0.85$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

It shows that the horizon was never reached and the majority of runs stop dynamic programming before 20 iterations.

## 4.5 Second Experiment

In this experiment we were considering both preferred states and preferred actions. Specifically in this experiment our preferred states was  $\mathcal{S}^i = \{5, 6\}$  and  $\mathcal{A}^i = \{3\}$ . Other input parameters were as follows  $\mu = 1$ ,  $w = 0.05$ ,  $|\mathcal{T}| = 50$ ,  $q^i = 0.85$  and length of the simulation was 300.

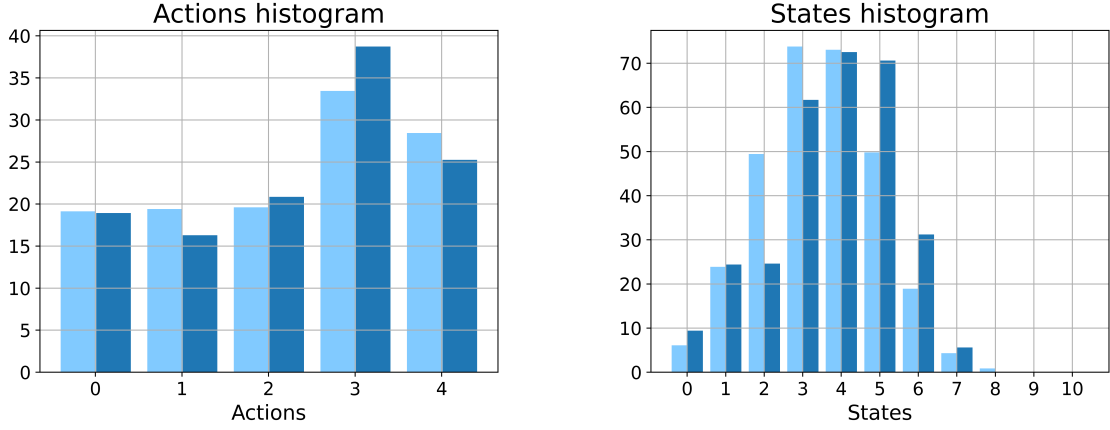
It is obvious, that evaluating the performance metrics in this experiment is more difficult, because in this experiment unlike the first experiment there is selected preferred action as well. The preferences on actions might contradict the preferences put on states. The weight  $w$  balances them.

The histograms of actions and states of the second experiment are shown below.

In Figure 4.3a we can see that for both methods action 3 was most used as we would expect with respect to PE. The method with fixed horizon used action 3 in more than one third of times, 103 time to be more precise. Our designed method with dynamic horizon used action 3 87 times.

In Figure 4.3b it can be seen that neither of both methods did especially well. The use of standard method with fixed horizon lead to observation of 71 times of state 5 and 30 times 6, which is around one third of all observed states. The use of our proposed method with dynamic horizon lead to observation of 58 times of state 5 and 18 times of state 6. That is drop in the realized preferred states by 25%.

Now we take closer look at the computational time performance of the second experiment.



(a) Histogram of actions in the experiment with preferred action  $\mathcal{A}^i = \{3\}$ .

(b) Histogram of states in the experiment with preferred states  $\mathcal{S}^i = \{5, 6\}$ .

Figure 4.3: Experiment with preferred states  $\mathcal{S}^i = \{5, 6\}$  and with preferred action  $\mathcal{A}^i = \{3\}$ . The input settings were  $\mu = 1$ ,  $w = 0.05$ ,  $q^i = 0.85$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

Type of Horizon	Median	Mean	Std. Deviation	Max.	Min.
Fixed	40.39	40.54	0.41	41.44	40.15
Dynamic	21.92	21.65	1.22	24.40	19.29

Table 4.2: Statistics of computational time (in seconds) for a single run of the experiment with the input settings  $\mathcal{S}^i = \{5, 6\}$ ,  $\mathcal{A}^i = \{3\}$ ,  $\mu = 1$ ,  $w = 0.05$ ,  $q^i = 0.85$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

In Table 4.2 we can see that in this experiment our proposed method of dynamic horizon lead to saving of computational time by around 45%, which is comparable with results in Table 4.1 from the previous experiment.

Last we show the histogram of the stopping time in the second experiment.

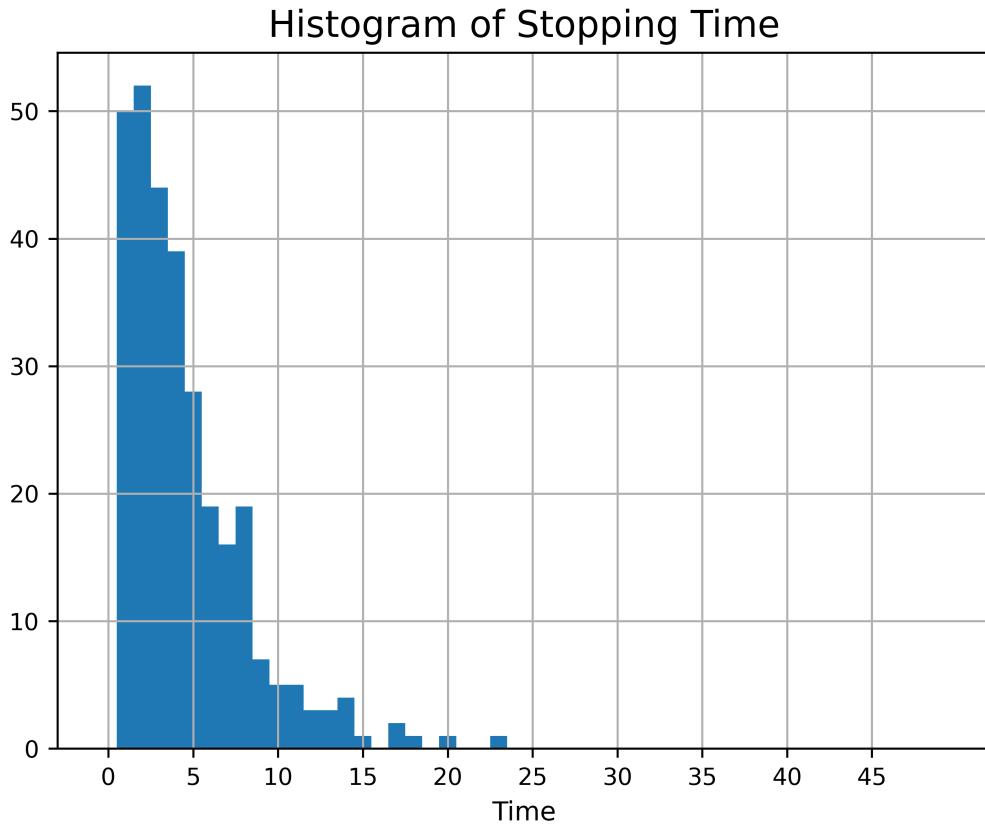


Figure 4.4: Histogram of stopping time of the experiment using input settings  $\mathcal{S}^i = \{5, 6\}$ ,  $\mathcal{A}^i = \{3\}$ ,  $\mu = 1$ ,  $w = 0.05$ ,  $q^i = 0.85$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

In Figure 4.4 we can see that histogram of stopping time resembles the exponential distribution and again the vast majority of the number of dynamic programming steps is below 20.

## 4.6 Third Experiment

In this experiment we focus on studying of influence of values  $q^i$  representing ideal probability of continuation, one of the key parameters in selection of dynamic horizon. We wanted to observe the influence of this parameter on computational time, stopping time and overall performance, such as the number of preferred states and actions realized. Other input settings are fixed at  $\mathcal{S}^i = \{3\}$ ,  $\mathcal{A}^i = \{1, 2\}$ ,  $\mu = 1$ ,  $w = 0.1$  and horizon  $|\mathcal{T}| = 50$ .

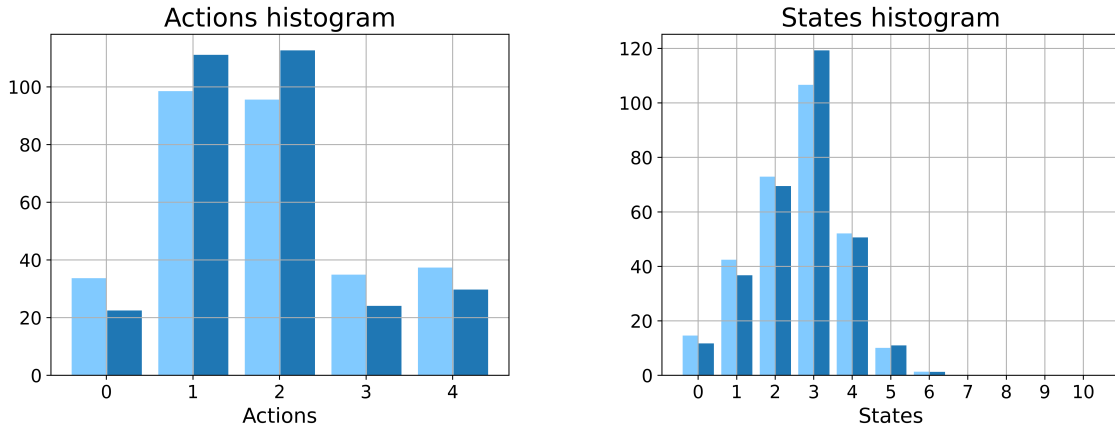
For the comparison of computational time for varying value of  $q^i$  see Table below.

Type of Horizon	$q^i$	Median	Mean	Std. Deviation	Max.	Min.
Fixed	–	37.62	37.87	0.85	41.44	37.54
Dynamic	0.85	19.70	19.43	1.23	21.73	16.94
Dynamic	0.90	25.40	25.09	1.83	28.22	22.19
Dynamic	0.95	40.73	39.81	3.31	44.73	32.86

Table 4.3: Statistics of computational time (in seconds) for a single run of the experiment with the input settings  $\mathcal{S}^i = \{3\}$ ,  $\mathcal{A}^i = \{1, 2\}$ ,  $\mu = 1$ ,  $w = 0.1$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs. The values of varying  $q^i$  are captured in the table.

In Table 4.3 it can be seen steadily increasing trend in computational time with increasing  $q^i$ , which is understandable. We can see that for  $q^i = 0.95$  computational time in median and in mean got over the computational time of method with fixed horizon as the stopping brings computational overheads.

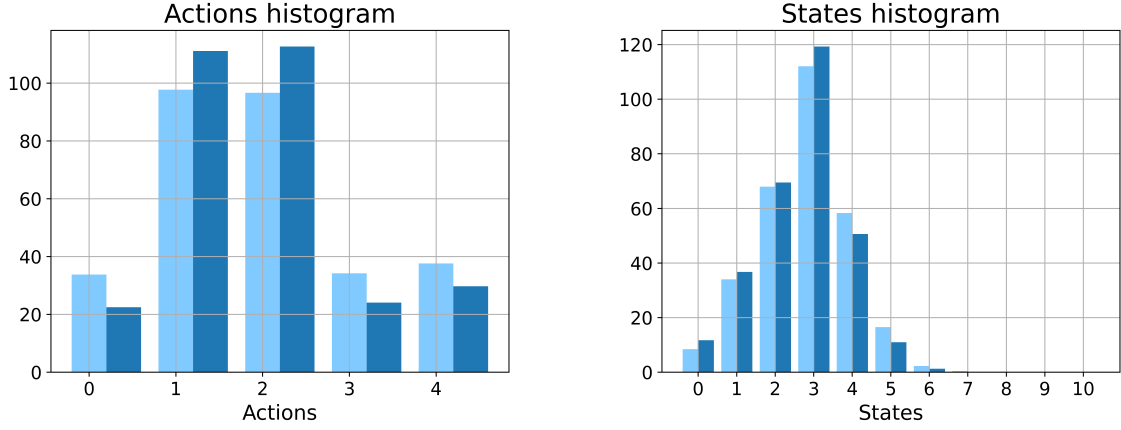
Now we are take a look at how did the histograms of actions and states changed based on the selected values  $q^i$ .



(a) Histogram of actions in the experiment with preferred action  $\mathcal{A}^i = \{1, 2\}$ . (b) Histogram of states in the experiment with preferred state  $\mathcal{S}^i = \{3\}$ .

Figure 4.5: Experiment with preferred state  $\mathcal{S}^i = \{3\}$  and the preferred actions  $\mathcal{A}^i = \{1, 2\}$ . The input settings were  $\mu = 1$ ,  $w = 0.1$ ,  $q^i = 0.85$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

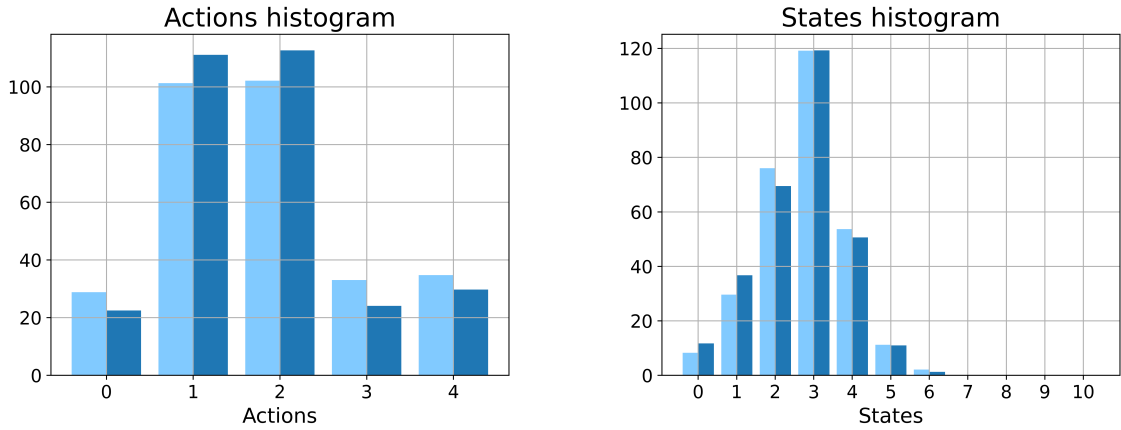




(a) Histogram of actions in the experiment with preferred action  $\mathcal{A}^i = \{1, 2\}$ .

(b) Histogram of states in the experiment with preferred state  $\mathcal{S}^i = \{3\}$ .

Figure 4.6: Experiment with preferred state  $\mathcal{S}^i = \{3\}$  and preferred actions  $\mathcal{A}^i = \{1, 2\}$ . The input settings were  $\mu = 1$ ,  $w = 0.1$ ,  $\mathbf{q}^i = \mathbf{0.90}$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.



(a) Histogram of states in experiment with preferred state  $\mathcal{S}^i = \{3\}$ .

(b) Histogram of states in experiment with preferred state  $\mathcal{S}^i = \{3\}$ .

Figure 4.7: Experiment with preferred state  $\mathcal{S}^i = \{3\}$  and preferred actions  $\mathcal{A}^i = \{1, 2\}$ . The input settings were  $\mu = 1$ ,  $w = 0.1$ ,  $\mathbf{q}^i = \mathbf{0.95}$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

From Figures 4.5a, 4.6a and 4.7a we can see that histogram of states observed when the dynamic horizon method was used were getting closer and closer to histogram of states generated using the standard fixed horizon method. That development makes sense and behaves as we could have expected. When  $\mathbf{q}^i = 1$  the method with dynamic horizon becomes the method with fixed horizon. In Figure 4.7b we can see, that with the value of parameter  $\mathbf{q}^i = 0.95$  counts of preferred state  $\mathcal{S}^i = \{3\}$  in dynamic scenario got almost same number as in standard method, that was a bit surprising and probably was caused by randomness in simulation. However this result suggests that with  $\mathbf{q}^i = 0.95$  observed results reaches comparable results with method with fixed horizon. This increase in the performance of the number of observed preferred states lead to the increase in the computational time, surpassing computational time of the method with fixed horizon.

However development of the generated actions did not change in the same way as the states did and

we can not see any regular dependence. Only fact that we can see from the displayed figures is that our proposed dynamic method does not put the same relevance or importance on the preferred actions as the standard method with fixed horizon.

Finally we are going to check the development of the stopping time based on varying  $q^i$ .

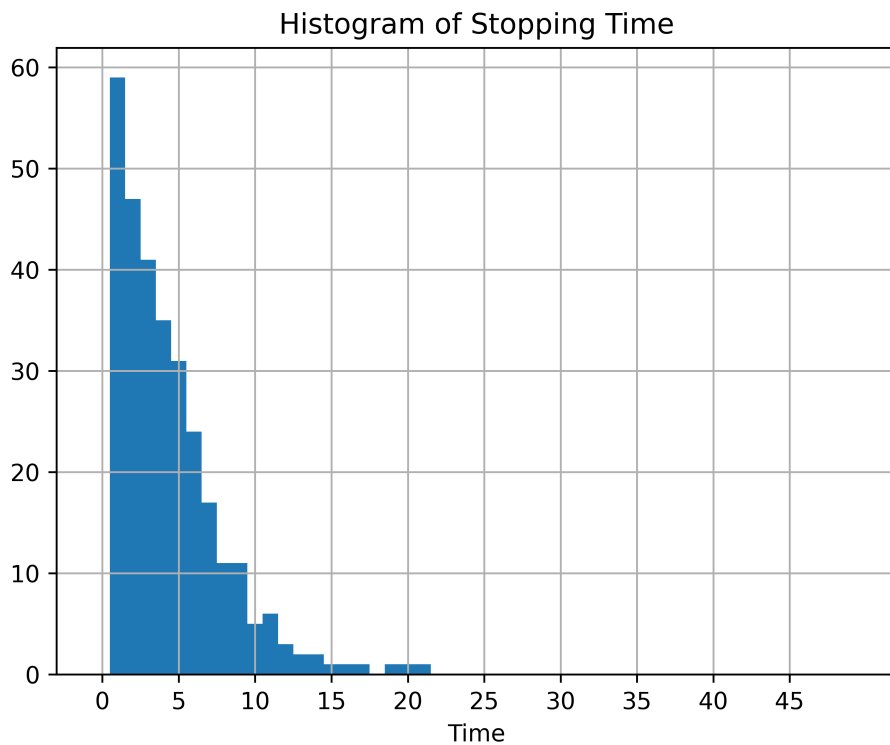


Figure 4.8: Histogram of stopping time of the experiment using input settings  $\mathcal{S}^i = \{3\}$ ,  $\mathcal{A}^i = \{1, 2\}$ ,  $\mu = 1$ ,  $w = 0.1$ ,  $q^i = \mathbf{0.85}$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

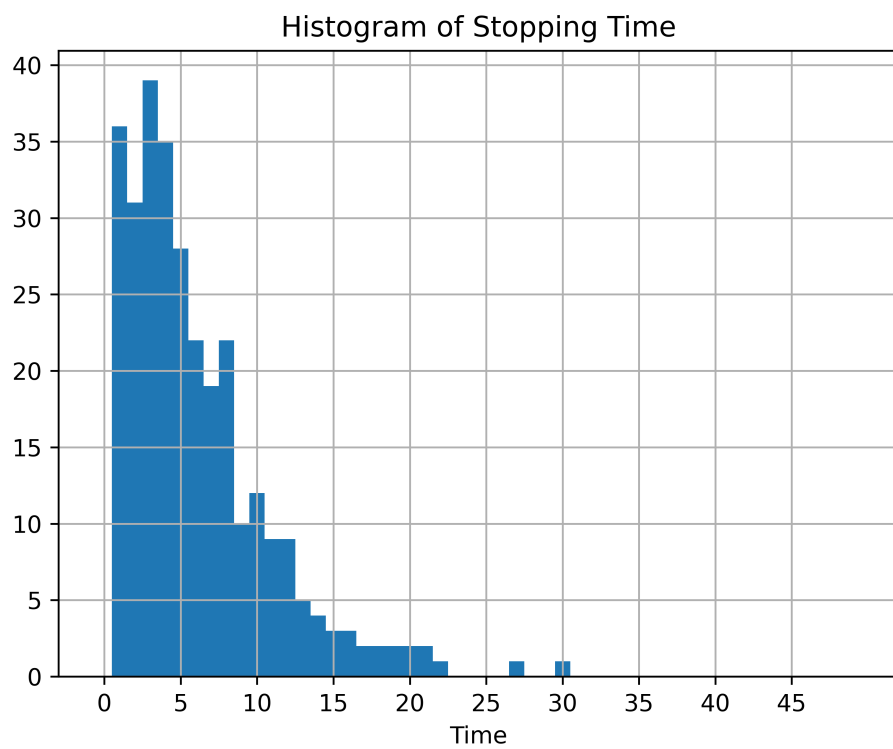


Figure 4.9: Histogram of stopping time of the experiment using input settings  $\mathcal{S}^i = \{3\}$ ,  $\mathcal{A}^i = \{1, 2\}$ ,  $\mu = 1$ ,  $w = 0.1$ ,  $\mathbf{q}^i = \mathbf{0.90}$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

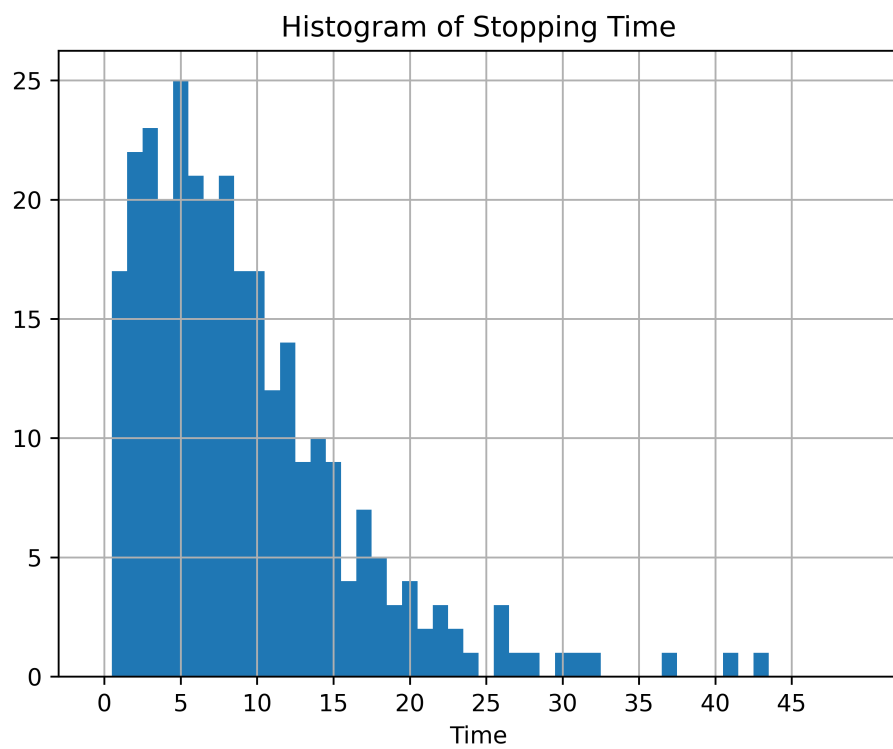


Figure 4.10: Histogram of stopping time of the experiment using input settings  $\mathcal{S}^i = \{3\}$ ,  $\mathcal{A}^i = \{1, 2\}$ ,  $\mu = 1$ ,  $w = 0.1$ ,  $\mathbf{q}^i = \mathbf{0.95}$ , horizon  $|\mathcal{T}| = 50$ , length of the simulation 300, made from 50 repeated runs.

In the Figures 4.8, 4.9 and 4.10 it is displayed increasing trend in stopping time as we expected. The stopping time in the experiment with  $q^i = 0.95$  our dynamic method reached peak in histogram around stopping time equal to 5. Displayed histograms are not that smooth, which is due to the insufficient number of repetitions of experiment. Higher number of repetition would solve this problem, but it was not within our technical capabilities to provide such a large number of independent evaluations.

As we can see from this experiment with selected values of  $q^i$  with higher values we got better results in form of number of observed preferred states, but for the price of increasing computational time. When with higher values of  $q^i$  getting closer to value 1, benefits of our proposed method in form of faster computational time blurs away.

## 4.7 Discussion

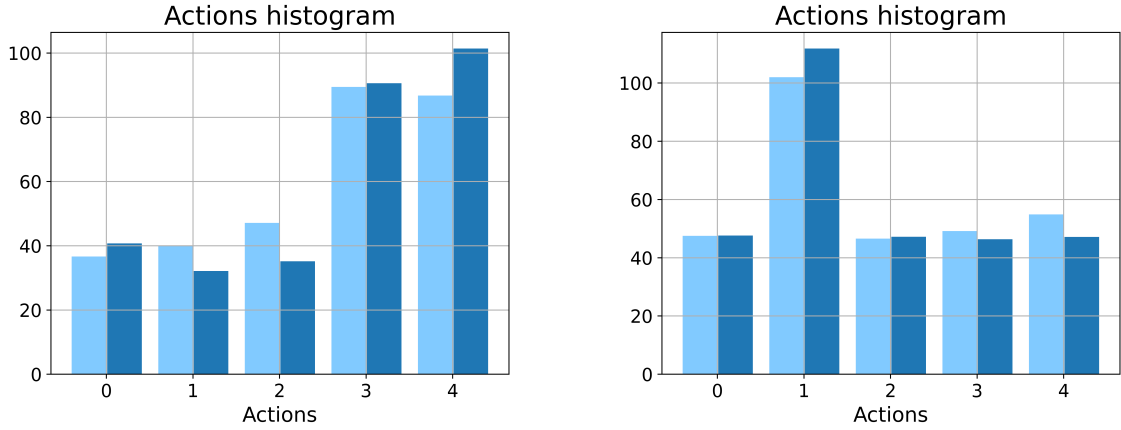
In this section, we summarize experience from all the previously presented simulations that are representative examples, including relevant details about the simulations that we tested but not presented in this work.

Our proposed method did relatively well in the first experiment, it lead to reasonable decrease in number of observed preferred states, but significant boost in computational time. Decrease in computational time was our main goal. At the same time, we wanted to maintain the quality of the decisions to achieve comparable results in meeting the preferences placed on states and actions. This goal was in this specific experiment satisfied.

In the second experiment we presented results, that was not so pleasant as the results from the first experiment, because of the drop in number of observed preferred states in comparison with standard method. Boost in computational time was still very good, around 45% in average. So maybe in this specific experiment it would be better to select higher value of  $q^i$ , to increase the reached quality with respect to observed states in our proposed dynamic method. But this decision what is more important, if time and computational costs or quality of observed states, is up to decision maker who is solving the specific problem.

The third experiment was focused on the parameter  $q^i$  and its influence in our proposed dynamic method. The presented results confirmed that with the increasing value of  $q^i$  the observed results of the dynamic method were getting closer to the standard method with fixed horizon. That trends seems to be reasonable to use.

When we take a look at the second and the third experiment it seems, that our proposed dynamic method cares less about the preferred actions, as in all results from these experiments our dynamic method used preferred actions less often than the standard method. We tried to briefly check this hypothesis on several other experiments with selected both preferred actions and states. They seem to support it. The samples of histograms of actions are displayed below, with specific settings of each experiment described are in Figure 4.11.



(a) Histogram of actions in experiment with preferred state  $\mathcal{S}^i = \{5\}$  and preferred actions  $\mathcal{A}^i = \{3, 4\}$ . The input settings were  $\mu = 1$ ,  $w = 0.05$ ,  $q^i = 0.90$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300.

(b) Histogram of actions in experiment with preferred state  $\mathcal{S}^i = \{6\}$  and preferred action  $\mathcal{A}^i = \{1\}$ . The input settings were  $\mu = 1$ ,  $w = 0.05$ ,  $q^i = 0.85$ ,  $|\mathcal{T}| = 50$  and the length of the simulation was 300.

Figure 4.11: Histograms of generated actions in selected experiments. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

The choice of the value  $|\mathcal{T}|$  is also worth discussion. The critical reader might suggest that all previously mentioned experiments were done on the horizon with value  $|\mathcal{T}| = 50$  and the stopping occurred much earlier. How would the proposed dynamic method be doing in the experiment with smaller horizon? These type of experiment and results are commented here, first from the computational time perspective and then from the quality of performance perspective. We show two examples how does both methods performed in cases with  $|\mathcal{T}| = 25$  and  $|\mathcal{T}| = 35$ .

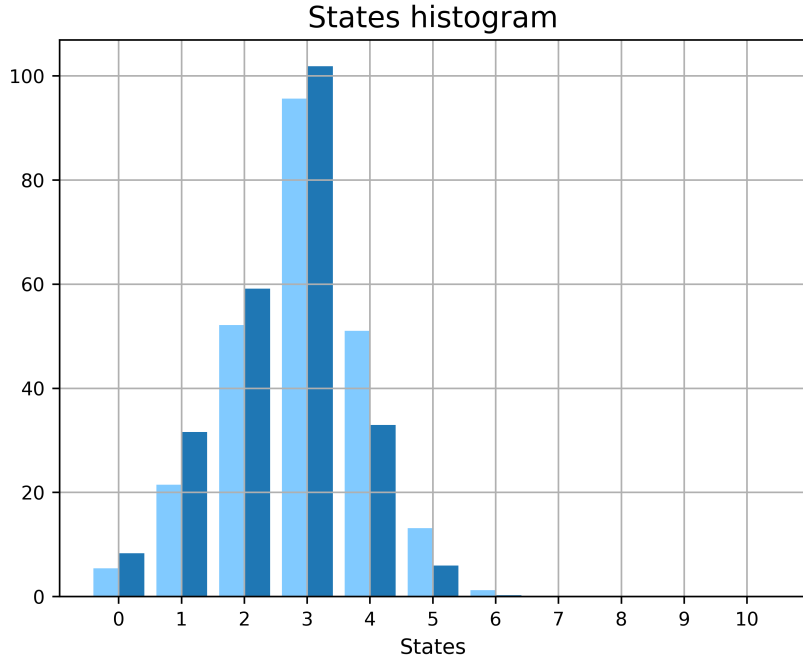
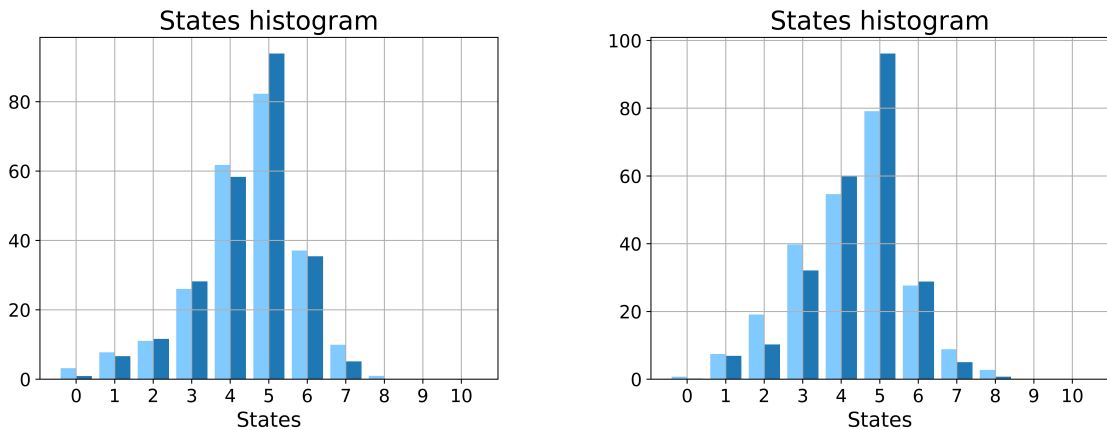


Figure 4.12: Histogram of states in experiment with preferred state  $\mathcal{S}^i = \{3\}$  and preferred actions  $\mathcal{A}^i = \{1, 2\}$ . The input settings were  $\mu = 1, w = 0.1, q^i = 0.80, |\mathcal{T}| = 25$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

Comparing Figure 4.12 with the results of the third example, namely, Figures 4.5b, 4.6b and 4.7b, it can be seen that drop in performance of both standard and dynamic method due to the selection of insufficiently large horizon.

Now we are going to present results from the experiment with not yet presented settings, showing its histograms of states.



(a) Histogram of states in the experiment with  $|\mathcal{T}| = 25$ . (b) Histogram of states in the experiment with  $|\mathcal{T}| = 35$ .

Figure 4.13: Histograms of generated states in experiments with different  $|\mathcal{T}|$ . The input settings of the experiment were  $\mathcal{S}^i = \{5\}, \mathcal{A}^i = \{3, 4\}, \mu = 1, w = 0.1, q^i = 0.90$  and the length of the simulation was 300. Counts for the method with fixed horizon with no stopping are darker, counts for the method with dynamic horizon with stopping are lighter.

In Figure 4.13 we can see significant increase in quality of decisions, due to the increase of the used horizon. It also confirms that the used value of horizon equal to  $|\mathcal{T}| = 50$  was sufficiently large. We checked that larger  $|\mathcal{T}|$  did not provide any noticeable improvement in the quality of the made decisions.

Finally in following table we present computational times of previously displayed experiments.

Type of Horizon	$ \mathcal{T} $	Median	Mean	Std. Deviation	Max.	Min.
Fixed	25	23.31	22.94	1.42	25.51	20.99
Dynamic	25	28.57	28.84	2.33	33.76	24.19
Fixed	35	32.15	32.00	2.07	37.69	29.33
Dynamic	35	30.31	30.08	2.78	35.21	24.66

Table 4.4: Statistics of computational time (in seconds) for a single run of the experiments with different values of horizon with the input settings  $\mathcal{S}^i = \{5\}$ ,  $\mathcal{A}^i = \{3, 4\}$ ,  $\mu = 1$ ,  $w = 0.1$ ,  $q^i = 0.90$ , length of the simulation 300, made from 50 repeated runs. The values of varying  $|\mathcal{T}|$  are captured in the table.

Finally, it is worth stressing, that when the selected preferences are unreasonable, then the standard design can not work very well and do not provide any good results. Similarly, our proposed method can not work well under these conditions.

Examples of such unreasonable preferences consisted of preferring states with transition probabilities close to zero (in our system, the states 9, 10) or selecting two separated groups of preferred states like  $\mathcal{S}^i = \{1, 2, 5, 6\}$ .

## Summary of Discussion

- Our proposed dynamic horizon method significantly improves computational time, often with a reasonable compromise in the achieved decision quality.
- The increasing the value of  $q^i$  made our dynamic method's results closer with the standard method with a fixed horizon.
- The dynamic method appeared less restrictive towards preferred actions in comparison to the standard method.
- The experiments with shorter horizons demonstrated that insufficiently large horizons deteriorate the performance quality of both methods.
- The unreasonable preferences cannot be met both the proposed dynamic method and the standard method.



## Chapter 5

# Conclusion

In this work, we have studied dynamic decision making with stopping using the FPD theory. The incorporation of the stopping into FPD is our main non-trivial contribution. FPD closed-loop formulation has a wide range of applications and any generalization of it will extend the applicability range.

Dynamic DM processes with stopping are very important part of the sequential DM area. As a possible examples of problems, which can be solved by our proposed FPD with stopping, is the famous Secretary Problem or any pharmaceutical testing of a new drug. In both those examples, the main task is to take a decision to stop in a right time, not too early, not too late. Our proposed solution can be used for evaluation of the optimal policy on all sorts of problem containing some form of the stopping, if can be expressed in the terms of closed-loop DM with stopping.

In the work, firstly, we have introduced the reader to the theory and notation necessary for the sufficient understanding of this work. Presented theory contained basics of probability theory, Markov Decision Process, Fully Probabilistic Design of Decision Strategies and Preference Elicitation.

Then, we have presented our proposition of extending the action and state space by stopping states and stopping actions. Those stopping states and actions represent indicators whether the DM process continues or is already stopped. These extended action and state spaces were after used for formulating the solution of DM processes with stopping using the extended form of FPD. This solution was formulated in a new theorem used for evaluation of the optimal policy in closed-loop process with stopping. This solution was then applied in the combination with preference elicitation (PE), which relieved necessity to design part of the ideal pds. The only necessary value to be designed, in addition to the values related to PE, is the ideal pd of continuation in dynamic programming, denoted  $q^i$  throughout this work. The outcome of this solution is the optimal rule, which then is used by agent to make a decision (generate action) to influence the system into the way agent prefers.

To demonstrate usability of our proposed method of solution in practice, we have implemented and performed a range of simulation experiments. In these experiments, to mimic the real-world problem, we have used Bayesian parametric estimation to iteratively update the estimates of the model of the unknown system. We have studied the impact of our proposed dynamic solution on the overall quality of the decisions taken and the impact on the time cost. We have compared our dynamic method with the standard method using standard FPD with PE. Our proposed method performed relatively well and in a number of experiments it provided a significant decrease in computation time at the cost of reasonable decrease in quality of taken decisions.

We see these results promising and worth of an effort for further improvements.

A more detailed and deeper study of selection the value of horizon  $|\mathcal{T}|$  and value  $q^i$  which are now inserted into the solution at the beginning based on prior expert knowledge, is needed. Designing a solution that automatically selects these values could lead to a more robust and general method.

Another possible extension of this work could be to use PE in the form of a dialogue with the user, rather than using time-invariant preferences as we did. This is because the user's preferences may change

during the DM task itself, and the dialogue can catch up with this development. This approach to PE was discussed in [8].

A direct comparison between our proposed approach and other dynamic methods on several different examples would be very useful for any future implementation. Experimental testing of the proposed solution on some real-world data could also help to identify advantages and limitations in real-world performance.

The most promising seems to be reformulation of FPD with PE and *partial stopping*. For instance, switching off estimation can spare computation time and prevent overfitting in a various models.

# Bibliography

- [1] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, 1978.
- [2] J. O. Berger. Bayesian analysis. *Statistical Decision Theory and Bayesian Analysis*, pages 118–307, 1985.
- [3] L. Chen and P. Pu. Survey of preference elicitation methods. Technical Report IC/2004/67, HCI Group Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2004.
- [4] J. S. Dyer, P. C. Fishburn, and et al. Multiple criteria decision making, multiattribute utility theory: The next ten years. *Man. Sci.*, 5(38):645–654, 1992.
- [5] T. S. Ferguson. Who solve the secretary problem? *Statistical Science*, 4(3):282–296, 1989.
- [6] M. Hoz. Fully probabilistic design of decision strategy with stopping. Bachelor’s thesis, FNSPE CTU, 2021.
- [7] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- [8] M. Kárný and T. Siváková. Agent’s feedback in preference elicitation. In *20th Int. Conf. on Ubiquitous Computing and Communications, IUCC*, pages 421–429, 2021.
- [9] M. Kárný and T. V. Guy. Fully probabilistic control design. *Systems & Control Letters*, 55(4):259–265, 2006.
- [10] M. Kárný and T. V. Guy. On the origins of imperfection and apparent non-rationality. *Studies in Computational Intelligence*, 538:57–92, 2014.
- [11] M. Kárný and T. V. Guy. Preference elicitation within framework of fully probabilistic design of decision strategies. *Workshop on Adaptive and Learning Control Systems*, 52:239–244, 2019.
- [12] T. J. Lorenzen. Optimal stopping with sampling cost the secretary problem. *The Annals of Probability*, 9(1):167–172, 1981.
- [13] V. Peterka. Bayesian approach to system identification. *Trends & Progress in System Identification*, pages 239–304, 1981.
- [14] M. Puterman. *Markov Decision Processes*. John Wiley & Sons, 1994.
- [15] A. Rényi. *Teorie pravděpodobnosti*. Academia, 1972.
- [16] J. Rust. Dynamic programming. *The New Palgrave Dictionary of Economics*, pages 3133–3158, 2018.
- [17] I. Vajda. *Teorie informace*. Vydavatelství ČVUT, 2004.
- [18] M. Virius. *Metoda Monte Carlo*. Vydavatelství ČVUT, 2010.

- [19] A. Wald. *Sequential Analysis*. Dover Publications, 2013.
- [20] L. Zwaan and H. Singh. The challenges in defining and measuring diagnostic error. *Diagnosis*, 2(2):97–103, 2015.