# A Unified Framework for Resource-Bounded Autonomous Agents Interacting with Unknown Environments

Pedro A. Ortega
Department of Engineering
University of Cambridge

A thesis submitted for the degree of
*Doctor of Philosophy*

September 2010

## Important!

I'm very proud to announce that this thesis has been downloaded over 150 times from my homepage (as of November 2014), which is something I did not expect back when I wrote it. Thank you!

Please be aware that this thesis is currently being updated to include the latest findings and exciting new chapters (bounded-rational decision trees, games, dynamic decision making, ...). If you have any comments and/or suggestions on how to improve this manuscript, please contact me at `pedro.ortega@gmail.com`.

*To my parents Pedro and Pilar.*

# Acknowledgements

This thesis is the result of four years of work, and it would not have been possible without the motivation and support of many people. Thanks to them, the years I spent in Cambridge have been amongst the happiest of my life.

I especially want to thank my family Pedro, Pilar, Carolina and Paulina and my closest friends Francisca Albert, Paul Aguayo, Daniel Braun, Oscar Van Heerden, Emre Karaa, Aliff Mohamad, José Donoso, Aditya Saxena, Loreto Valenzuela, Horacio Tate, Aaron Lobo, Zoi Roupakia, Aysha Roohi, Ben Mansfield, Francisco Pérez and Mauricio Gaete. Also, I am deeply indebted to Disa Helander and Raffaella Nativio.

Special thanks go to my friend Daniel Braun, who has been closely collaborating with me during the course of my study. I also want to thank José Aliste, John Cunningham, José Donoso, Marcus Hutter, Humberto Maturana, Gonzalo Ruz and David Wingate for their invaluable help and comments on earlier versions of this manuscript. The present study has been supported by the Ministerio de Planificación de Chile (MIDEPLAN), the Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) and the Böhringer-Ingelheim-Fonds (BIF). Finally, I want to thank my supervisor Zoubin Ghahramani for his guidance, motivation and support.

# Abstract

The aim of this thesis is to present a mathematical framework for conceptualizing and constructing adaptive autonomous systems under resource constraints. The first part of this thesis contains a concise presentation of the foundations of classical agency: namely the formalizations of decision making and learning. Decision making includes: (a) subjective expected utility (SEU) theory, the framework of decision making under uncertainty; (b) the maximum SEU principle to choose the optimal solution; and (c) its application to the design of autonomous systems, culminating in the Bellman optimality equations. Learning includes: (a) Bayesian probability theory, the theory for reasoning under uncertainty that extends logic; and (b) Bayes-Optimal agents, the application of Bayesian probability theory to the design of optimal adaptive agents. Then, two major problems of the maximum SEU principle are highlighted: (a) the prohibitive computational costs and (b) the need for the causal precedence of the choice of the policy. The second part of this thesis tackles the two aforementioned problems. First, an information-theoretic notion of resources in autonomous systems is established. Second, a framework for resource-bounded agency is introduced. This includes: (a) a maximum bounded SEU principle that is derived from a set of axioms of utility; (b) an axiomatic model of probabilistic causality, which is applied for the formalization of autonomous systems having uncertainty over their policy and environment; and (c) the Bayesian control rule, which is derived from the maximum bounded SEU principle and the model of causality, implementing a stochastic adaptive control law that deals with the case where autonomous agents are uncertain about their policy and environment.

# Contents

## CONTENTS

# List of Figures

## LIST OF FIGURES

# List of Notation

## Basic

| | |
|---|---|
| $\mathcal{X}$ | A set or alphabet. |
| $\mathbb{N}$ | The set of natural numbers $1, 2, 3, \ldots$ |
| $\mathbb{R}$ | The set of real numbers. |
| $\epsilon$ | The empty string. |
| $\mathcal{X}^n$ | The set of strings of length $n$ over the alphabet $\mathcal{X}$. |
| $\mathcal{X}^*$ | The set of all finite strings over the alphabet $\mathcal{X}$. |
| $x_{i:k}$ | The substring $x_i x_{i+1} \cdots x_{k-1} x_k$. |
| $x_{\leq i}$ | The string $x_1 x_2 \cdots x_i$. |
| $\ln(x)$ | The natural logarithm of $x$. |
| $\log(x)$ | The logarithm base-2 of $x$. |
| $\mathscr{P}(\mathcal{X})$ | The powerset of $\mathcal{X}$. |
| $\mathbf{Pr}$ | An arbitrary probability distribution. |
| $\Omega$ | A sample space. |
| $\mathcal{F}$ | An algebra. |

## Autonomous Agents

| | |
|---|---|
| $\mathcal{A}$ | The set of actions. |
| $\mathcal{O}$ | The set of observations. |
| $\mathcal{Z}$ | The set of interactions, i.e. $\mathcal{Z} := \mathcal{A} \times \mathcal{O}$. |
| $a_t$ | The action at time $t$. |
| $o_t$ | The observation at time $t$. |
| $\underline{ao}_i$ | An interaction viewed as a symbol, i.e. $a_i o_i$. |
| $T$ | The horizon, i.e. the maximum length of interaction strings. |
| $\mathcal{Z}^\diamond$ | The set of interaction strings up to length $T$. |
| $\mathbf{U}$ | The utility function over interactions strings. |
| $\mathbf{P}$ | The (behavioral) model of the agent, i.e. the distribution over interaction sequences implemented by the agent. It is used to denote the input model, the output model or the I/O model. |

| | |
|---|---|
| $\mathbf{Q}$ | The (behavioral) model of the environment, i.e. the distribution over interaction sequences implemented by the environment. It is used to denote the input model, the output model the I/O model. |
| $\mathbf{G}$ | The generative distribution, i.e. the sampling distribution over the interactions sequences that results from the interaction between the agent and the environment. |
| $P$ | The belief model of the agent, i.e. the Bayesian mixture distribution over interaction sequences. The symbol is used to denote the Bayesian input model, the Bayesian output model or the Bayesian I/O model. The belief model is a conceptual explanation that gives rise to a unique behavioral model. |

## Model Types

| | |
|---|---|
| $\mathbf{P}(a_t \| \underline{ao}_{<t})$ | The probability of generating action $a_t$ given the past $\underline{ao}_{<t}$. The collection of these probabilities forms the agent's output model. |
| $\mathbf{P}(o_t \| \underline{ao}_{<t} a_t)$ | The probability of gathering observation $o_t$ given the past $\underline{ao}_{<t} a_t$. The collection of these probabilities forms the agent's input model. |
| " | The collection of the previous two forms the agent's I/O model. |
| $\mathbf{Q}(a_t \| \underline{ao}_{<t})$ | The probability of gathering action action $a_t$ given the past $\underline{ao}_{<t}$. The collection of these probabilities forms the environment's input model. |
| $\mathbf{Q}(o_t \| \underline{ao}_{<t} a_t)$ | The probability of generating observation $o_t$ given the past $\underline{ao}_{<t} a_t$. The collection of these probabilities forms the environment's output model. |
| " | The collection of the previous two forms the environment's I/O model. |
| $P(\theta)$ | The prior probability of the parameter $\theta \in \Theta$. |
| $P(a_t \| \theta, \underline{ao}_{<t})$ | The probability of generating action $a_t$ given the past $\underline{ao}_{<t}$ under the parameter $\theta$. Together with the prior over $\theta$, the collection of these probabilities forms the agent's Bayesian output model. |
| $P(o_t \| \theta, \underline{ao}_{<t} a_t)$ | The probability of gathering observation $o_t$ given the past $\underline{ao}_{<t} a_t$ under the parameter $\theta$. Together with the prior over $\theta$, the collection of these probabilities forms the agent's Bayesian input model. |
| " | The collection of the previous three forms the agent's Bayesian I/O model. |

## Resources & Boundedness

| | |
|---|---|
| $\gamma$ | The conversion factor between units of information and units of energy. |
| $\alpha$ | The conversion factor between units of information and units of utility. |
| $\mathbf{U}(A)$ | The utility of event $A$. |
| $\mathbf{u}(A\|B)$ | The utility gain of changing from event $B$ to event $A \cap B$, i.e. $\mathbf{u}(A\|B) := \mathbf{U}(A \cap B) - \mathbf{U}(B)$. |
| $\mathbf{J}(\mathbf{Pr}; \mathbf{U})$ | The free utility of $\mathbf{Pr}$ under the utility $\mathbf{U}$. |
| $\mathbf{P}_0$ | In a transformation, this is the distribution that is assumed to be known (It can be the distribution before or the distribution after the transformation). |
| $\mathbf{Pr}$ | In a variational problem, this is the distribution to be varied. |
| $\mathbf{P}_i, \mathbf{S}$ | In a transformation, this is the distribution before the change. |
| $\mathbf{P}_f, \mathbf{R}$ | In a transformation, this is the distribution after the change. |

# LIST OF NOTATION

# List of Definitions and Results

## LIST OF DEFINITIONS AND RESULTS

# Preface

Artificial intelligence is a fascinating field. It is the only field of engineering that deals with your most familiar possession—your mind! How do we *create* intelligence? And, by the way, *what is it*?

When I first read Marcus Hutter's *Universal Artificial Intelligence* (2004a) I was astonished—it is the optimal solution to the artificial intelligence problem! The AIXI agent defined therein optimally adapts to any (computable) situation: in particular, it will make predictions about the weather, play chess, solve mazes, drive cars, discover physical theories, solve IQ tests and predict future stock prices. Even though, admittedly, AIXI cannot be implemented because it is uncomputable, one should still be able to devise a down-scaled approach to create primitive intelligence. Or maybe not?

Well, down-scaling turns out to be a *very* difficult task. The literature is rich of (implicit) "approximations" to the optimal solution, and probably there is at least one new method being conceived every week. Many of these approximations resort to ad-hoc methods that deviate considerably from the theory of rationality, and hence one ends up asking oneself whether they can be justified at all from the point of view of the theory. Moreover, it is a fact that all these state-of-the-art approximations either only work in very restricted problem domains or, if they aim to be more general, cannot cope with more than simple toy-problems. The computational complexity of these approximations are just way too high. This is disappointing.

However, spiders build webs, ants and bees successfully navigate complex terrains to collect food. Simple organisms display intelligence levels that seem inexplicable given their limited resources—at least from the point of view of the theory of rationality.

Rather than starting a search for a "better" approximation, and driven by my dampened optimism about the theory, I wanted to find out whether there is more to rationality than we actually know. This necessarily implies investigating the well-established foundations of decision making and learning that we nowadays take for granted. It turns out that one can point out at least two fundamental shortcomings: the causal precedence of the choice of the policy and the implicit computational bottleneck, both of them imposed by the very theory of rational decision making!

Hence, my research program for the past four years has been to understand where the limitations imposed by the theory of rationality arise, and to formulate a framework of bounded rationality that includes classical rationality as a limit case.

The goal of this thesis is to present the foundations of classical rationality along with

a synthesis of my work during my graduate studies at the Department of Engineering of the University of Cambridge. Most of the effort in writing this thesis has been allocated in presenting the material and results in the most concise and simplest way I could (hopefully, with moderate success), especially because the fundamental ideas are very simple! In other words: one brief and coherent story, no fancy mathematics, and to drop (admittedly with sadness) the topics that do not add anything substantial to the main message.

## Structure of this Thesis

The thesis contains ten short chapters and is divided into two parts: *Foundations of Classical Agency* and *Resource-Bounded Agency*. The two parts also happen to roughly subdivide the material into the preexisting literature and the original contributions, respectively.

**Chapter 1: Introduction.** This chapter introduces the reader to the problem of the design of autonomous agents.

**Part I: Foundations of Classical Agency.**

**Chapter 2: Characterization of Behavior.** The aim of this chapter is to familiarize the reader with the basic notation & concepts that are necessary in order to characterize behavior. In particular, it clarifies what is meant by a model of an autonomous system, and introduces the first (and simplest) of the four models of this thesis.

**Chapter 3: Decision Making.** The principle of maximum subjective expected utility is the standard design principle for the construction of cybernetic systems. Where does it come from and what is its justification? This chapter presents the underlying theory.

**Chapter 4: Learning.** When the cybernetic system faces an unknown environment, then it has to adapt to it by learning through experience. How is this learning process formalized? This chapter reviews the standard framework for reasoning under uncertainty, and applies it to the case of cybernetic systems.

**Chapter 5: Problems of Classical Agency.** This chapter highlights the main drawbacks of classical agency.

**Part II: Resource-Bounded Agency.**

**Chapter 6: Resources.** What are resources and how are they formalized? This chapter provides an information-theoretic answer to this question.

**Chapter 7: Boundedness.** Based on three intuitive axioms, one can derive a conversion law between utilities and probabilities (and hence, information or resources). This conversion law is then used to formulate a bounded subjective expected utility that models decision-making under resource constraints.

**Chapter 8: Causality.** When a cybernetic system gathers inputs or generates outputs, its information state is updated. However, inputs and outputs have different updates. While the first is well modeled by the learning framework, the latter requires the consideration of causal constraints. This chapter introduces a framework to deal with both inputs and outputs.

**Chapter 9: Control as Estimation.** This chapter combines the theory developed in the two preceding chapters to formulate a new law for adaptive control. Furthermore, a convergence result for a very restricted case is provided.

**Chapter 10: Discussion.** This chapter briefly discusses the results and concludes.

## Declaration of Originality

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

# Chapter 1

# Introduction

In artificial intelligence and control theory, a major field of ongoing research is the design of complex autonomous systems[1] interacting with unknown environments. Good designs are not easy to achieve because they must consider and integrate solutions to many different problems. Apart from the purely physical sensory-motor challenges in the case of embodied agents, a well-designed agent must cope with several information processing problems at the same time. From these, learning, decision making, and the limited amount of (computational) resources play the major roles.

The first formal treatment of autonomous systems can be traced back to the ideas of the field of *cybernetics* in the first half of the twentieth century (Rosenblueth, Wiener, and Bigelow, 1943). Since then, and especially after the development of cheap and fast computers during the decade of 1980, artificial intelligence and control theory have evolved into rich and fruitful areas of research. Nowadays, the ideas that have emerged from both fields are present everywhere, ranging from applications in large-scale industrial manufacturing to small embedded systems and single-user software.

The theoretical foundations for the design of optimal autonomous systems for a given problem class is well-developed and widely accepted. *It is even well understood how to characterize optimal universal autonomous systems that can adapt to any (computable[2]) environment* (Hutter, 2004a). However, the theory has not yet led to real-world implementations of complex autonomous systems. Even though recently promising approximations have been developed (Veness, Ng, Hutter, and Silver, 2010; Veness, Ng, Hutter, W., and Silver, 2011), the problem remains essentially unresolved in practice. The main reason for this failure is simple: *the computational complexity of constructing such a system is prohibitive.*

The aim of this thesis is to review the theoretical foundations of agency, to identify its shortcomings, and to propose a mathematical formalization of resource-bounded agency. This is a necessary enterprise given the current state-of-the-art: virtually all

[1] Originally, *autonomous systems* were called *cybernetic systems.*

[2] The term *computable* refers to functions that are the formalized analogue of the intuitive notion of an algorithm. The Church-Turing thesis postulates that computable functions are exactly the functions that can be calculated using a mechanical calculation device given unlimited amounts of time and storage space.

of the research community centers its efforts around the design of increasingly more efficient algorithms that are approximations to the "gold standard" dictated by the theory. There is no consensus, however, about how to implement, how to rationalize, nor how to compare such approximations.

## 1.1 Historical Remarks & References

*Cybernetics* was pioneered by Rosenblueth et al. (1943) and then established as a field by Wiener's book *Cybernetics* (Wiener, 1965). The field later evolved to what is now known as *control theory*, where purposeful behavior is conceived as the minimization of the "error" between the current state and the goal state. This field, mainly driven by electrical engineers, developed a rich theory of continuous, mainly linear, control systems. Nevertheless, the mathematics used in control theory also imposed limitations to the kind of autonomous systems that could be conceived. As a response to this, J. McCarthy, M. Minsky, C. Shannon, A. Turing, R. Solomonoff, and other researchers gave birth to the field of *artificial intelligence* (AI)—"the science and engineering of making intelligent machines". Inspired by the work of McCulloch and Pitts (1943), AI researchers embraced the idea of duplicating human faculties like creativity and self-improvement.

Between the 1950s and 1970s, AI grew into a rich and fruitful field with many very diverse approaches, mainly centered around solving restricted problem domains (a paradigm now called *narrow AI*). Later, during the 1980s, the rapid development of cheap and fast computers paved the way for the adoption of advanced statistical and control-theoretic techniques, thus revitalizing the interest for the problem of the design of complex autonomous systems. Perhaps the most influential modern books, presenting many of the disparate and isolated ideas in AI and related fields in a unified way, are Russell and Norvig (2009) and Nilsson (1998). Another modern approach that has considerably grown in popularity is *reinforcement learning* (Sutton and Barto, 1998). In this setup, an autonomous system learns *what to do* from a feedback signal (in addition to the observation signal) issued by the environment. The system system uses this feedback signal to evaluate its performance. Within the reinforcement learning paradigm, *universal artificial intelligence* has established itself as a rigorous top-down approach to AI (Hutter, 2004a; Legg, 2008), combining ideas from sequential decision making with universal prediction (Solomonoff, 1964). This work presents an optimal (though uncomputable) autonomous system (called AIXI) that is universal with respect to the class of computable environments.

# Part I

# Foundations of Classical Agency

# Chapter 2

# Characterization of Behavior

An **autonomous system** is anything that has an *observable input and output* (I/O) stream, like a calculator, a human cell, an animal, a computer program or a robot. In other words, it is anything having a boundary defining what is inside and what is outside and acting as an interface to communicate with its environment. This definition is symmetrical: the environment of an autonomous system is an autonomous system too.

A autonomous system communicates with its environment by interacting with it. This interaction consists in the exchange of *symbols*, generated in order to influence each other. The generation of these symbols is governed by (stochastic) rules that make up the *behavior* of the autonomous system.

There are certain behaviors that are more desirable than others. Artificial intelligence and control theory can be seen as the design and analysis of the behavior of autonomous systems. The goal of this chapter is to introduce mathematical formalizations of behavior that should serve as a basis for the study of systems and their interactions.

## 2.1 Preliminaries

We first establish some notation, conventions and basic concepts.

### 2.1.1 Basic Notation

A **set** is denoted by a calligraphic letter like $\mathcal{X}$ and consists of **elements** or **symbols**. We use $\mathbb{N} = \{1, 2, 3, \ldots\}$ for the set of **natural numbers**, and $\mathbb{R}$ for the set of **real numbers**. **Strings** are finite concatenations of symbols. The **empty string** is denoted by $\epsilon$. $\mathcal{X}^n$ denotes the set of strings of length $n$ based on $\mathcal{X}$. For **substrings**, the following shorthand notation is used: a string that runs from index $i$ to $k$ is written as $x_{i:k} := x_i x_{i+1} \ldots x_{k-1} x_k$. Similarly, $x_{\leq i} := x_1 x_2 \ldots x_i$ is a string starting from the first index. By convention, $x_{i:j} := \epsilon$ if $i > j$. **Logarithms** are always taken with respect to base 2, thus $\log(2) = 1$, unless written explicitly $\ln(x)$, in which case we mean **natural**

**logarithms**. The symbol $\mathscr{P}(\mathcal{X})$ denotes the **powerset** of $\mathcal{X}$, i.e. the set of all subsets of $\mathcal{X}$.

### 2.1.2 Probabilities & Random Variables

To simplify the exposition, all probability spaces are assumed to be finite unless clearly stated otherwise. While this assumption limits the domain of application of the exposition, it allows isolating the problems belonging solely to the design of autonomous systems from the problems that arise due to infinite sets (and in particular, from the problems of geometrical or topological assumptions). Due to this, we clarify some terminology.

A **sample space** is a finite set $\Omega$, where each member $\omega \in \Omega$ is called a **sample** or **outcome**. A subset $A \subset \Omega$ is called an **event**. A subset $\mathcal{F} \subset \mathscr{P}(\Omega)$ of events that contains $\Omega$ and is closed under complementation and finite union is called an **algebra**. A **measurable space** is a tuple $(\Omega, \mathcal{F})$, where $\Omega$ is a sample space and $\mathcal{F}$ is an algebra. Given a measurable space $(\Omega, \mathcal{F})$, a set function $\mathbf{Pr}$ over $\mathcal{F}$ is called a **probability measure** iff it obeys the (Kolmogorov) **probability axioms**

K1. for all $A \in \mathcal{F}$, $\quad \mathbf{Pr}(A) \in [0, 1]$;

K2. $\mathbf{Pr}(\varnothing) = 0$ and $\mathbf{Pr}(\Omega) = 1$;

K3. for all disjoint $A, B \in \mathcal{F}$, $\quad \mathbf{Pr}(A \cup B) = \mathbf{Pr}(A) + \mathbf{Pr}(B)$.

A **probability space** is a tuple $(\Omega, \mathcal{F}, \mathbf{Pr})$ where $(\Omega, \mathcal{F})$ is a measurable space and $\mathbf{Pr}$ is its measure. Given a probability space $(\Omega, \mathcal{F}, \mathbf{Pr})$, a **random variable** is a function $X : \Omega \to \mathcal{X}$ mapping each outcome $\omega$ into a symbol $X(\omega)$ of a set $\mathcal{X}$, and where $X^{-1}(x) \in \mathcal{F}$ for all $x \in \mathcal{X}$. The probability of the random variable $X$ taking on the value $x \in \mathcal{X}$ is defined as $\mathbf{Pr}(x) := \mathbf{Pr}(X = x) := \mathbf{Pr}(\{\omega \in \Omega : X(\omega) = x\})$.

## 2.2 Models of Autonomous Systems

If one wants to characterize the way an autonomous system behaves, it is necessary to fully describe the rules governing its potential I/O stream. The mathematical description of an autonomous system's behavior can be done at several levels, ranging from very detailed physical descriptions (e.g. in the case of a robot, it would require specifying its gears, distribution of current, sensors and effectors, etc.) to abstract statistical descriptions. This thesis deals exclusively with statistical descriptions. In particular, during the course of this thesis, two levels of description will be discussed: namely, the behavioral level and the belief level.

1. *Behavioral Level.* The behavioral level specifies the actual I/O behavior, i.e. the probabilities of making observations and issuing actions of the autonomous system. These probabilities have to be specified for every possible information state of the system, meaning that they must characterize the system's I/O statistics

for every possible sequence of past I/O symbols. In Chapter 6, we will link this description with the amount of resources (measured in thermodynamic work) that the autonomous system has to spend during its interactions. The behavioral level will be introduced in two parts: first, by specifying the output model and then by completing it with an input model.

   (a) *Output Model.* The output model specifies how the autonomous system generates its outputs given the past I/O symbols. This is the minimal statistical description of an autonomous system's behavior. However, it does not explain the purpose of the system, i.e. it does not explain what it tries to accomplish. This model will be introduced in this chapter.

   (b) *Input Model.* The autonomous system predicts its input stream using the input model. This prediction model represents the assumptions that the system makes about its environment. We will see in Chapter 3 that these assumptions are necessary in order to formalize the purpose of the autonomous system. Furthermore, we will argue in Chapter 6 that this model is also necessary in order to characterize the thermodynamical work spent in interactions. The input model is introduced in the next chapter.

2. *Belief Level.* In this thesis we are mainly interested in modeling *adaptive* autonomous agents. However, specifying adaptive agents by directly writing down their behavioral models is difficult. To simplify the description of adaptive agents, one first starts out specifying a belief model, which is a high-level auxiliary model characterizing the beliefs, assumptions and uncertainties of an agent. Subsequently, one simply derives the low-level behavioral model from the belief model. This is similar to writing programs in a high-level programming language that is then compiled into low-level machine code. As in the case of the behavioral level, this belief level too will be introduced in two parts: first, by introducing the Bayesian input model and then by introducing the Bayesian output model.

   (a) *Bayesian Input Model.* When the autonomous system does not know its environment, then the designer can use a Bayesian approach to model this uncertainty. This Bayesian input model is currently the standard in the adaptive control literature. It is introduced in Chapter 4.

   (b) *Bayesian Output Model.* In addition to modeling the uncertainty an autonomous system has over its environment, a designer can also model the uncertainty the system has over its own policy. This leads to a Bayesian output model that is analogous to the Bayesian input model. However, we will see that this generalization is not straightforward, as it will require a careful distinction between the technical treatment of actions and observation arising due to causal constraints. This model is an original contribution of this thesis, and it will be introduced in Chapter 8, Part II.

A schematic illustration of this setup is given in Figure 2.1. Other aspects that are important in the characterization of autonomous systems are the following:

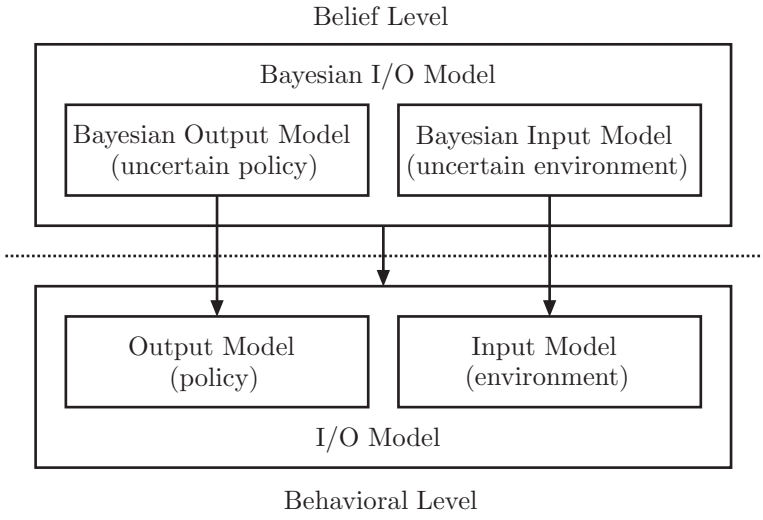



**Figure 2.1:** Behavioral Models. Dependencies are indicated by arrows.

1. *I/O Domains:* The choice of the cardinality, geometry and topology of the I/O domain can have a significant impact on the description, the implementation and the performance of an autonomous system. In navigation devices for instance, the usage of continuous I/O sets is vital for its robustness. Furthermore, choosing and modeling continuous I/O sets appropriately can be a very challenging task. However, in this thesis we will exclusively deal with finite I/O sets, because arguably the core problems of the design of autonomous agents are already present even in this simple setup.

2. *Interaction Protocol:* The interaction protocol specifies all the details concerning the rules and the timing of the communication between two autonomous systems. There are applications where de-synchronized interactions with variable timing play a crucial role, like e.g. in moving artillery or in tennis. In this thesis it is assumed that interactions occur in discrete time steps where autonomous systems alternately take turns to generate a symbol. While in many cases, this interaction protocol is flexible enough to accommodate other interaction regimes (e.g. simultaneous interactions, discrete approximations of continuous time, etc.), it is not clear whether the theory would significantly change under other interaction protocols.

3. *Multiple Autonomous Systems:* This thesis deals exclusively with situations having only two interacting autonomous systems. However, there are many situations, especially in economics, where one would prefer modeling the behavior of a population of autonomous systems. If the population is significantly large, then a more coarse-grained description of behavior could be beneficial: e.g. treating

entire sub-populations as if they were individual autonomous systems; or using a completely different description modeling emergent properties. This view might be especially relevant to the understanding of decentralized behavior for instance.

## 2.3 Output Model

In the following an autonomous system's behavior is formalized as a conditional probability measure over I/O sequences over I/O alphabets. We start with basic definitions of interactions.

**Definition 1 (Interactions)** The possible I/O symbols are drawn from two finite sets. Let $\mathcal{O}$ denote the set of **observations** and let $\mathcal{A}$ denote the set of **actions**. The set $\mathcal{Z} := \mathcal{A} \times \mathcal{O}$ is the set of **interactions**. The interaction string of length 0 is denoted by $\epsilon$. Let $T \in \mathbb{N}$ be the **horizon**, i.e. maximum length of interaction strings. Let $\mathcal{Z}^\diamond := \bigcup_{t=0}^{T} \mathcal{Z}^t$ denote the set of interactions up to length $T$. We also underline symbols to glue them together as for example in $\underline{ao}_{\leq 2} = a_1 o_1 a_2 o_2$. □

We assume that there are two autonomous systems $\mathbf{P}$ and $\mathbf{Q}$. By convention, we assume that $\mathbf{P}$ is the **agent** i.e. the autonomous system to be constructed by the designer, and that $\mathbf{Q}$ is the **environment**[1], i.e. the autonomous system to be controlled by the agent.

Agent and environment operate obeying the following interaction protocol. The interaction proceeds in **cycles** $t = 1, 2, \ldots, T$. In cycle $t$, the agent $\mathbf{P}$ generates an action $a_t$ conditioned on the past I/O symbols $\underline{ao}_{<t}$. The action is observed by both systems. Then, the environment $\mathbf{Q}$ responds by generating an observation $o_t$ conditioned on the past $\underline{ao}_{<t} a_t$. The observation is observed by both systems as well. Then, the next cycle starts. This interaction protocol is fairly general, and many other interaction protocols can be translated into this scheme. Figure 2.2 illustrates this setup.

The definitions of the agent's and the environment's output model follow.

**Definition 2 (Output Model)** An **output model** of an agent is a set of conditional probabilities

$$\mathbf{P}(a_t | \underline{ao}_{<t}), \qquad \text{for all} \quad \underline{ao}_{\leq t} \in \mathcal{Z}^\diamond,$$

inducing a unique probability measure $\mathbf{P}$ over $\mathcal{A}^T$ conditioned on $\mathcal{O}^T$ given by

$$\mathbf{P}(a_{\leq t} | o_{<t}) := \prod_{\tau=1}^{t} \mathbf{P}(a_\tau | \underline{ao}_{<\tau}).$$

Similarly, an **output model** of an environment is a set of conditional probabilities

$$\mathbf{Q}(o_t | \underline{ao}_{<t} a_t), \qquad \text{for all} \quad \underline{ao}_{\leq t} \in \mathcal{Z}^\diamond,$$

[1] *Agent* and *environment* are the terms commonly used in the artificial intelligence literature. In the control literature, the two autonomous systems are known as *controller* and *plant* respectively.

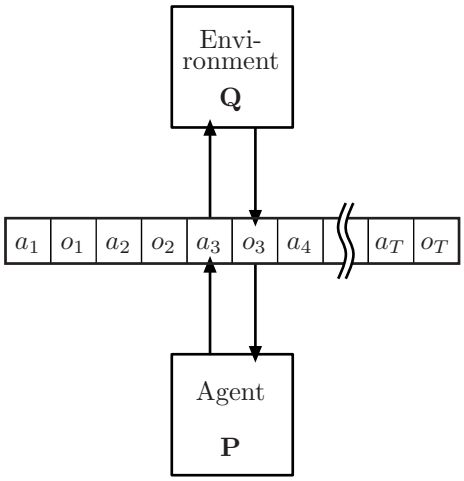



**Figure 2.2:** An interaction system. The agent $\mathbf{P}$ and the environment $\mathbf{Q}$ define a probability distribution over interaction sequences $\mathcal{Z}^\diamond$.

inducing a unique probability measure $\mathbf{Q}$ over $\mathcal{O}^T$ conditioned on $\mathcal{A}^T$ given by

$$\mathbf{Q}(o_{\leq t}|a_{\leq t}) := \prod_{\tau=1}^{t} \mathbf{Q}(o_\tau|\underline{ao}_{<\tau}a_\tau).$$

□

Thus, the output model for an agent is a probability distribution over the next action given the past, and the output model for an environment is a probability distribution over the next observation given the past. This type of specifications are called **stream probabilities**. The asymmetry in the definitions should not distract the reader from the fact that these choices are purely conventional. Output models are graphically represented as trees (Figure 2.3).

When an agent and an environment are coupled, they uniquely define a probability measure over the I/O strings $\mathcal{Z}^\diamond$, i.e. the probability law governing the actual production of I/O symbols. This probability measure is given by the generative probability distribution defined next.

**Definition 3 (Generative Probability Measure)** Let $\mathbf{P}$ and $\mathbf{Q}$ be the output model of an agent and an environment respectively. The **generative probability measure** $\mathbf{G}$ over $\mathcal{Z}^\diamond$ is defined by the conditional probabilities

$$\mathbf{G}(a_t|\underline{ao}_{<t}) := \mathbf{P}(a_t|\underline{ao}_{<t})$$
$$\mathbf{G}(o_t|\underline{ao}_{<t}a_t) := \mathbf{Q}(o_t|\underline{ao}_{<t}a_t)$$

valid for all $\underline{ao}_{\leq t} \in \mathcal{Z}^\diamond$. □

The generative probability measure $\mathbf{G}$ is the objective probability law from which the interaction sequences are sampled. As such, it is the probability law that the

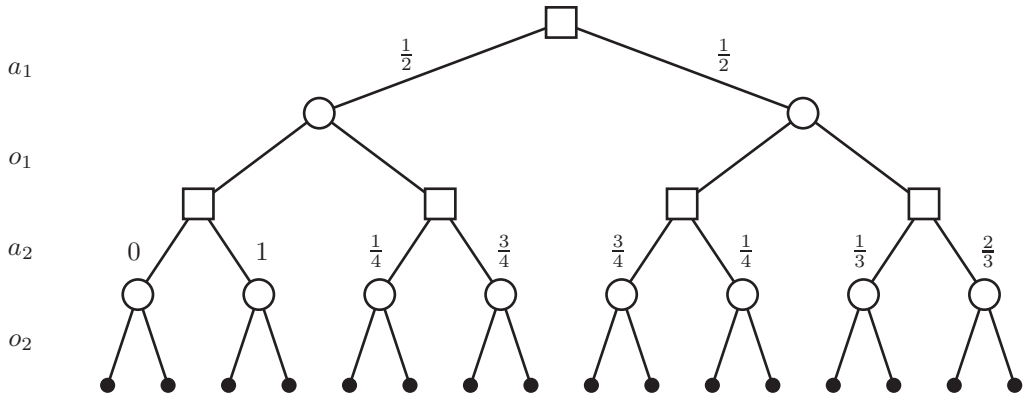



**Figure 2.3:** An Output Model. Behavioral models are naturally represented as trees. The figure illustrates an output model for an agent (i.e. a distribution over actions given past I/O symbols) for binary I/O sets $\mathcal{A} := \mathcal{O} := \{0, 1\}$. In the tree there are three types of nodes. Square nodes (□) are action nodes, round nodes (◯) are observation nodes and leaves (•) are full I/O histories. A transition from a parent node to a child node on the left corresponds to choosing symbol 0, and a transition to a child node on the right corresponds to choosing symbol 1. In this case, only the transition probabilities in the action nodes are specified because observations are just conditionals. The probability of choosing symbol $a_t$ in the node reached by following the path $\underline{ao}_{<t}$ starting from the root node is given by $\mathbf{P}(a_t|\underline{ao}_{<t})$. For instance, $\mathbf{P}(a_2 = 1|a_1 = 0, o_1 = 1) = \frac{3}{4}$.

designer should use in order to assess the objective expected performance of a system. However, in practice this probability distribution is unknown because the environment is unknown. In this case, the designer can assume a "subjective" probability measure over the I/O stream, i.e. a probability law that is thought to hold, and then use it to compute "subjective" expectations. This will be clarified in the next chapter.

# Chapter 3

# Decision Making

The output model introduced in the previous chapter contains all the information to characterize an autonomous system's behavior. However, it is not very useful for explaining *what* it tries to accomplish, i.e. it does not provide any insight about its *purpose*. One would like to have a framework that allows *justifying* the choice of an autonomous system's behavior. The standard framework for conceptualizing purpose is the theory of *subjective expected utility* (SEU) developed by Savage (1954).

## 3.1 Subjective Expected Utility

The theory of SEU is a framework dealing with decision making under uncertainty, i.e. a situation where a choice does not uniquely determine the outcome. We can motivate this as follows. Suppose you are running late to get to your work. You have the option to either wait for the bus, or run to your work. These choices are called "acts" in SEU. However, the bus can be unreliable, and there is a possibility of the bus not being on time. These uncertain situations are called "states" in SEU. Depending on whether you decide to wait for the bus or run to your work and whether the bus is on time, there are three possible "outcomes": either you are on time but tired (because you decided to run); or you are on time and active (you waited for the bus and it arrived on time); or you are late and active (you waited for the bus but it came too late). This decision problem is summarized in Table 3.1.

| | | Act | |
|---|---|---|---|
| | | Wait for bus. | Run to work. |
| State | Bus is late. | Late and active. | On time and tired. |
| | Bus is on time. | On time and active. | On time and tired. |

**Table 3.1:** A Decision Problem with Uncertainty.

SEU theory postulates that rational decision makers can be characterized by their

preferences over acts (and mixtures over acts), and that these preferences obey certain structural and consistency rules. For instance, you might prefer to "wait for the bus" over "either waiting for the bus or running to work with equal probabilities". Intuitively, it seems plausible to say that this preference reveals something about both your subjective belief in the bus being on time and your subjective desire of getting to your job active. This is precisely the idea underlying SEU: if preferences obey certain rationality axioms, then these preferences reveal the subjective beliefs in events happening and the subjective desirability of the outcomes.

### 3.1.1 Setup

SEU theory is formalized as follows. Let $\mathcal{S}$ be a set of **states**, $\mathcal{C}$ be a set of **consequences** and $\mathcal{F}$ be a set of **acts** (that is, mappings from the set of states $\mathcal{S}$ into the set of consequences $\mathcal{C}$). This is illustrated in Figure 3.1. The intuitive meaning is as follows: an act is a choice available to the decision maker; a consequence is a possible result of an act; and a state is a compilation of facts about the world that the decision maker is uncertain about but uniquely determines the consequence of a chosen act. A subset of states $A \subset \mathcal{S}$ is an **event**. Note that the states of world $\mathcal{S}$ are mutually exclusive and complete, i.e. exactly one state $s \in \mathcal{S}$ will occur.

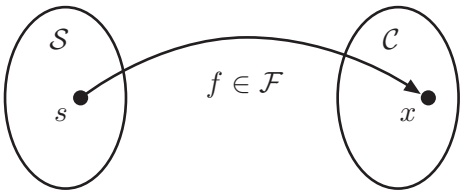



**Figure 3.1:** Setup of Subjective Expected Utility.

### 3.1.2 Rationality

A decision maker is characterized by a **preference relation** $\succcurlyeq$ on the set of acts $\mathcal{F}$, i.e. $\succcurlyeq$ is transitive and complete. For any pair of acts $f, g \in \mathcal{F}$, the expression $f \succcurlyeq g$ means that "the decision maker *prefers* $f$ over $g$". From $\succcurlyeq$, one constructs the **strict preference relation** $\succ$ and the **indifference relation** $\sim$ defined as

$$\begin{aligned} f \succ g \quad &\Leftrightarrow \quad (f \succcurlyeq g) \quad \text{and not} \quad (g \succcurlyeq f), \\ f \sim g \quad &\Leftrightarrow \quad (f \succcurlyeq g) \quad \text{and} \quad (g \succcurlyeq f), \end{aligned}$$

and having the obvious meaning. The decision maker is called **rational** if its preference relation $\succcurlyeq$ fulfills the axioms of rationality presented further down. Some definitions are needed before.

**Definition 4 (Conditional Preference)** $f$ is **preferred** over $g$ **given** $A \subset \mathcal{S}$, written $f \succcurlyeq_A g$, iff $f' \succcurlyeq g'$ where $f = f'$ and $g = g'$ on $A$ and $f' = g'$ on $A^c$. □

Intuitively, $f \succcurlyeq_A g$ means that $f$ is preferred over $g$ if we disregard all the states outside of $A$. The auxiliary functions $f'$ and $g'$ are constructed such that comparing $f'$ and $g'$ unconditionally will say only something about how $f$ and $g$ compare within $A$. This is because $f'$ and $g'$ are equal to $f$ and $g$ within $A$ respectively, but are indistinguishable from each other outside of $A$. The relations $\succ_A$ and $\sim_A$ are constructed in the obvious way from $\succcurlyeq_A$.

**Definition 5 (Null Event)** A subset $A \subset \mathcal{S}$ is said to be **null** iff for all $f, g \in \mathcal{F}$, $f$ is indifferent over $g$ given $A$, i.e. $f \sim_A g$. □

Null sets correspond to the events where the decision maker is indifferent between all of his available choices. They will turn out to be the events having zero probability.

**Definition 6 (Constant Act)** An act $f$ is said to be **constant** iff for all $s \in \mathcal{S}$, $f(s) = x \in \mathcal{C}$ for some $x \in \mathcal{C}$. In this case, we also simply write $x$ to denote the constant act $f$. □

Note that a constant act $x$ is a "sure gamble", because the consequence $x$ is the same irrespective of the state of the world $s$.

We first present the axioms of rationality together in one place and subsequently briefly discuss their meaning. The following version corresponds to Kreps' presentation of the axioms (Kreps, 1988), but they are essentially the same as the original ones introduced by Savage (1954).

**Definition 7 (Rationality/Savage Axioms)** Let $\mathcal{S}$, $\mathcal{C}$ and $\mathcal{F}$ be a set of states, a set of outcomes, and a set of acts respectively. A binary relation $\succcurlyeq$ over $\mathcal{F}$ is said to be a **rational preference relation** iff it follows the axioms:

R1. $\succcurlyeq$ is transitive and complete.

R2. There exist $x, y \in \mathcal{C}$ such that $x \succ y$.

R3. Let $f, f', g, g' \in \mathcal{F}$ and $A \subset \mathcal{S}$ be such that
(a) $f(s) = f'(s)$ and $g(s) = g'(s)$ if $s \in A$,
(b) $f(s) = g(s)$ and $f'(s) = g'(s)$ if $s \notin A$.
Then, $f \succ g$ iff $f' \succ g'$.

R4. Let $A \subset \mathcal{S}$ be non-null and for all $s \in A$, $f(s) = x$ and $g(s) = y$.
Then $f \succ_A g$ iff $x \succ y$.

R5. Let $x, y, x', y' \in \mathcal{C}$, $f, g, f', g' \in \mathcal{F}$, and $A, B \subset \mathcal{S}$ be such that
(a) $x \succ y$ and $x' \succ y'$,
(b) $f(s) = x$ and $f'(s) = x'$ for $s \in A$, and $f(s) = y$ and $f'(s) = y'$ for $s \in A^c$,
(c) $g(s) = x$ and $g'(s) = x'$ for $s \in B$, and $g(s) = y$ and $g'(s) = y'$ for $s \in B^c$.
Then, $f \succ g$ iff $f' \succ g'$.

R6. For all $A \subset \mathcal{S}$,
(a) $[f \succ_A g(s)$ for all $s \in A]$ implies $f \succcurlyeq_A g$,
(b) $[g(s) \succ_A f$ for all $s \in A]$ implies $g \succcurlyeq_A f$.

R7. For all $f, g \in \mathcal{F}$ such that $f \succ g$ and for all $x \in \mathcal{C}$,
there is a finite partition of $\mathcal{S}$ such that for every $A$ in the partition,
(a) $[f'(s) = x$ for $s \in A, f'(s) = f(s)$ for $s \in A^c]$ implies $f' \succ g$,
(b) $[g'(s) = x$ for $s \in A, g'(s) = g(s)$ for $s \in A^c]$ implies $f \succ g'$.

Axiom R1 just states that $\succcurlyeq$ is a preference relation.

Axiom R2 is purely structural: it rules out the trivial situation where the decision maker is indifferent between all the consequences.

Axiom R3 justifies talking about conditional preference. Consider the situation in Figure 3.2 and the comparison of $f$ with $g$. Essentially, Axiom R3 tells us that we only need to worry about how they compare in $A$, since they agree on $A^c$. Thus, if there is another pair $f'$ and $g'$ that agrees on $A^c$ and is identical to $f$ and $g$ respectively on $A$, then we can conclude the preference order of $f$ and $g$ from $f'$ and $g'$.

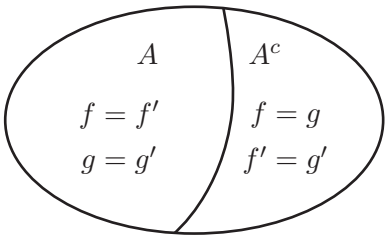



**Figure 3.2:** The setup in Axiom R3.

Axiom R4 tells us that utilities of outcomes are not state-dependent. That is, if the decision maker strictly prefers an outcome $x$ over $y$ (and thereby strictly prefers a constant act $x$ over $y$), then knowing that the (non-null) event $A$ obtains preserves the strict preference order.

Axiom R5 essentially tells us that acts cannot affect probabilities. Let the "prizes" (i.e. consequences) $x$ and $y$ be a "win" and a "loss" respectively. Consider two "gambles" (i.e. acts) $f$ and $g$ over these outcomes having different winning odds, represented by the different winning sets of states $A$ and $B$ respectively, as depicted in Figure 3.3 a. Then, if we prefer $f$ over $g$, then it is because the odds in $f$ are more favorable than in $g$. Hence, replacing the prizes $x$ and $y$ by $x'$ and $y'$ having the same order (i.e. $x' \succ y'$) will preserve the preference order over the resulting gambles (i.e. $f' \succ g'$, see Figure 3.3 b).

Axiom R6 is the *sure-thing principle*. Consider for instance the situation depicted in Figure 3.4, where the consequences in $\mathcal{C}$ are linearly ordered in increasing order of preference. The acts $f$ and $g$ map the possible states in $A$ into the sets of consequences

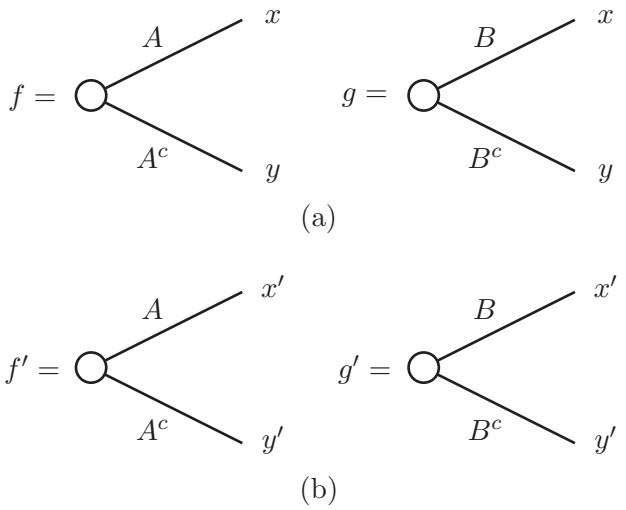



**Figure 3.3:** Example illustrating Axiom R5.

$f(A)$ and $g(A)$ respectively. If an act $f$ is strictly preferred over any "sure gamble" $g(s)$ with $s \in A$, then $f$ is preferred over any "randomizing gamble" $g$ over $A$ by the force of Axiom R6.

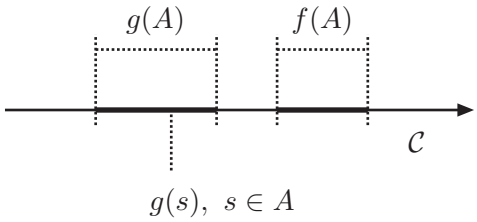



**Figure 3.4:** Example illustrating Axiom R6.

Axiom R7 is a "continuity" axiom. Consider two acts $f, g$ such that $f \succ g$ and an arbitrary constant act $x$. Then, by Axiom R7, there is always a (sufficiently fine-grained) finite partition of $\mathcal{S}$, such that all mappings $f'$ constructed from $f$ by replacing a small "patch" of the partition by the constant act $x$ (Figure 3.5) are still strictly preferred over $g$, i.e. slightly changing $f$ into $f'$ will still preserve the preference relation $f' \succ g$. Similarly, slightly changing $g$ into $g'$ will preserve the preference relation $f \succ g'$.

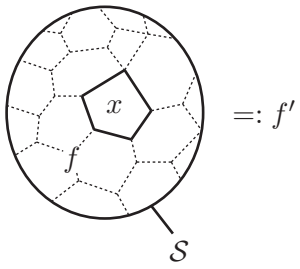



**Figure 3.5:** Example illustrating Axiom R7.

### 3.1.3 Representation Theorem

The main result of SEU is that any rational decision maker can be thought of as evaluating an expected utility. Formally:

**Theorem 1 (Expected Utility Representation Theorem)** *Let $\succcurlyeq$ be a preference relation on $\mathcal{F}$. Then the following two conditions are equivalent:*

*i. $\succcurlyeq$ satisfies axioms R1-R7.*

*ii. There exists a unique, nonatomic, finitely additive, probability measure $\mathbf{P}$ on $\mathcal{S}$ such that $\mathbf{P}(A) = 0$ iff $A$ is null, and a bounded, unique up to a positive affine transformation, real-valued function $\mathbf{U}$ on $\mathcal{C}$ such that, for all acts $f$ and $g$, $f \succcurlyeq g$ iff*

$$\int_{\mathcal{S}} \mathbf{U}(f(s))\, d\mathbf{P}(s) \geq \int_{\mathcal{S}} \mathbf{U}(g(s))\, d\mathbf{P}(s).$$

*In this expression, the notation $\int_{\mathcal{S}} \phi(\omega)\, d\mathbf{P}$ denotes the Lebesgue-integral of a function $\phi$ over the set of states $\mathcal{S}$ with respect to the measure $\mathbf{P}$.* □

**Remark 1** The rationality axioms R1–R7 imply that the underlying probability space is uncountably infinite. This is a common technical assumption that is used to obtain uniqueness results. For our purposes, we will assume that the utility function $\mathbf{U}$ and the probability measure $\mathbf{P}$ are given, and think about them conceptually as resulting from such a process of comparing acts. □

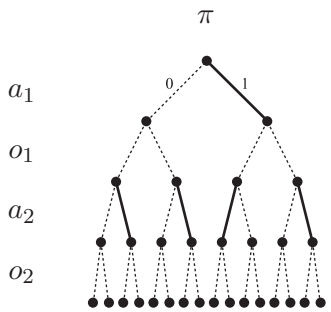
Put differently, if a decision maker's preference relation $\succcurlyeq$ is rational, then the theorem justifies the existence of two (essentially) unique *subjective entities*: a probability measure $\mathbf{P}$ representing the plausibilities of the states of the world, and a utility function $\mathbf{U}$ representing how desirable the consequences are. Moreover, the decision maker's choice behavior can then be formalized as maximizing the expected utility. Conversely, having a probability measure $\mathbf{P}$ and a utility function $\mathbf{U}$, and basing choices on the maximization of the expected utility, implies that the decision-maker has a rational preference relation $\succcurlyeq$.

## 3.2 The Maximum Subjective Expected Utility Principle

An autonomous agent is a decision maker. Therefore, equipped with the theory of SEU, one can construct the behavior of a *rational* autonomous agent by maximizing an expected utility. This leads to a construction method called the *maximum subjective expected utility principle* (abbreviated maximum SEU principle). But before we can apply SEU theory to interaction systems, we need to develop a formal correspondence between the language of SEU theory and the language of autonomous systems. This is the aim of the following section. While the concepts that we will present in the following are well-known in the literature, their formal correspondence to SEU theory is an original contribution of this thesis.

### 3.2.1 SEU in Autonomous Systems

In the SEU framework, an act is a deterministic mapping of a states of the world into consequences—there seems to be no randomness at all. This is not so: the randomness arises because the decision maker has imperfect knowledge about the state of world, represented by his subjective probability measure. Similarly, here we will assume that an autonomous system with random behavior can be represented as a collection of deterministic autonomous systems and a probability measure over them. If an autonomous system is deterministic, then it can be represented as a function mapping a past I/O string into an I/O symbol (Figure 3.6).




| π | |
|---|---|
| $\underline{ao}_{<t}$ | $\pi(\underline{ao}_{<t})$ |
| $\epsilon$ | 1 |
| 0 0 | 1 |
| 0 1 | 1 |
| 1 0 | 0 |
| 1 1 | 1 |

**Figure 3.6:** A behavioral function $\pi$ is most naturally represented as a transition tree. Interactions ($\mathcal{A} = \mathcal{O} = \{0, 1\}$) start from the root node. In the figure, choices are highlighted with solid edges. Note that $\pi$ only chooses transitions in the $a_t$-levels, not in the $o_t$-levels. The table specifies the function $\pi$.

**Definition 8 (Behavioral Function)** A **behavioral function** of an agent is a mapping $\pi : \mathcal{Z}^\diamond \to \mathcal{A}$ of histories $\underline{ao}_{<t}$ into actions $a_t = \pi(\underline{ao}_{<t})$. Similarly, a **behavioral function** of an environment is a mapping $\lambda : \mathcal{Z}^\diamond \times \mathcal{A} \to \mathcal{O}$ of histories $\underline{ao}_{<t}a_t$ into observations $o_t = \lambda(\underline{ao}_{<t}a_t)$. □

Let $\Pi$ and $\Lambda$ be the set of behavioral functions of agents and environments respectively. It is easily seen (Figure 3.7) that each choice of a pair $(\pi, \lambda) \in \Pi \times \Lambda$ determines a unique interaction string $\underline{a^* o^*}_{\leq T} \in \mathcal{Z}^T$ as

$$a_t^* = \pi(\underline{a^* o^*}_{<t}) \qquad \text{and} \qquad o_t^* = \lambda(\underline{a^* o^*}_{<t} a_t^*). \tag{3.1}$$

Put differently, choosing $(\pi, \lambda)$ also chooses a unique interaction string in $\mathcal{Z}^T$. For each $\pi \in \Pi$, define the agent functional $\phi_\pi : \Lambda \to \mathcal{Z}^T$ as the functional mapping a behavioral function of an environment $\lambda \in \Lambda$ into the interaction string $\phi_\pi(\lambda) \in \mathcal{Z}^T$ defined by (3.1). Let $\Phi := \{\phi_\pi | \pi \in \Pi\}$ be the set of these agent functionals.

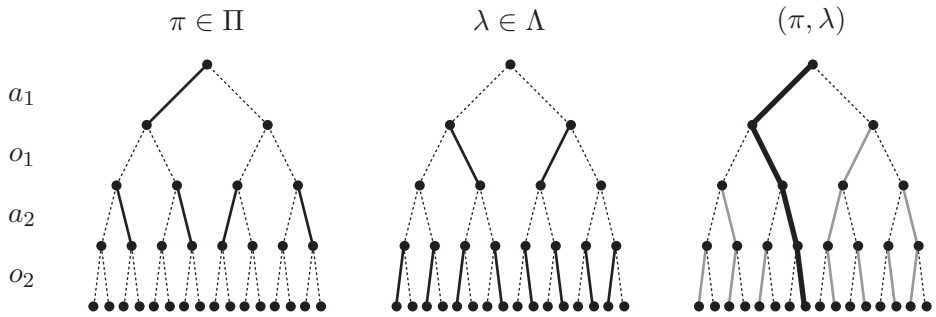



**Figure 3.7:** Choosing behavioral functions of an agent and of an environment determines a unique interaction string.

With these definitions, one can establish the following correspondence to SEU:

| | | | | |
|---|---|---|---|---|
| States | $\mathcal{S}$ | $\longleftrightarrow$ | $\Lambda$ | Beh. Func. of Environments |
| Consequences | $\mathcal{C}$ | $\longleftrightarrow$ | $\mathcal{Z}^T$ | Interaction Strings |
| Acts | $\mathcal{F}$ | $\longleftrightarrow$ | $\Phi$ | Agent Functionals |

Additionally, we assume that we are given a subjective probability measure $\mathbf{P}$ over $\Lambda$ and a subjective utility function $\mathbf{U}$ over $\mathcal{Z}^T$. This correspondence sets up a scheme where we can determine whether an agent functional $\phi_{\pi_1}$ is preferred over an agent functional $\phi_{\pi_2}$ by comparing their subjective expected utilities:

$$\phi_{\pi_1} \succcurlyeq \phi_{\pi_2} \qquad \Longleftrightarrow \qquad \sum_{\lambda \in \Lambda} \mathbf{P}(\lambda) \mathbf{U}(\phi_{\pi_1}(\lambda)) \geq \sum_{\lambda \in \Lambda} \mathbf{P}(\lambda) \mathbf{U}(\phi_{\pi_2}(\lambda)).$$

This in turn allows us extending the preference relation over agent functionals to a preference relation over behavioral functions of agents by defining

$$\pi_1 \succcurlyeq \pi_2 \qquad \Longleftrightarrow \qquad \phi_{\pi_1} \succcurlyeq \phi_{\pi_2}.$$

Finally, our last step of the correspondence links the previous notions of behavioral functions to stream probabilities. Because a behavioral function $\lambda \in \Lambda$ of an environment specifies how to choose the observation $o_t \in \mathcal{O}$ for any past $\underline{ao}_{<t} a_t$, having a

probability measure $\mathbf{P}$ over $\Lambda$ induces a set of conditional probabilities

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) := \sum_{\lambda\in\Lambda} \mathbf{P}(\lambda)\, \delta_\lambda(o_t|\underline{ao}_{<t}a_t),$$

where

$$\delta_\lambda(o_t|\underline{ao}_{<t}a_t) := \begin{cases} 1 & \text{if } o_t = \lambda(\underline{ao}_{<t}a_t), \\ 0 & \text{otherwise.} \end{cases}$$

In a similar fashion, choosing a behavioral function of an agent $\pi \in \Pi$ induces a set of (degenerate) conditional probabilities

$$\mathbf{P}(a_t|\underline{ao}_{<t}) := \begin{cases} 1 & \text{if } a_t = \pi(\underline{ao}_{<t}), \\ 0 & \text{otherwise.} \end{cases}$$

Note how we have defined a way of extending the probability measure $\mathbf{P}$ over $\Lambda$ to a probability measure over the interaction strings $\mathcal{Z}^\diamond$. Importantly, it is worth stressing that fixing $\pi \in \Pi$ corresponds to choosing an output model $\mathbf{P}(a_t|\underline{ao}_{<t})$ (Definition 2).

The overall conclusion is as follows. If an autonomous system is rational, then it chooses its output model $\mathbf{P}(a_t|\underline{ao}_{<t})$ such that is maximizes the expected utility

$$\sum_{\underline{ao}_{\leq T}} \mathbf{P}(\underline{ao}_{\leq T})\mathbf{U}(\underline{ao}_{\leq T}), \tag{3.2}$$

where the utility function $\mathbf{U}$ over $\mathcal{Z}^T$ and the set of conditional probabilities $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ are given. Note that the probability measure $\mathbf{P}$ over $\mathcal{Z}^T$ is obtained from the conditional probabilities $\mathbf{P}(a_t|\underline{ao}_{<t})$ and $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ using the product rule for probabilities. This construction method is the well-known **maximum SEU principle** for autonomous systems (Russell and Norvig, 2009).

**Remark 2** Solving (3.2) by choosing the optimal behavioral function $\pi \in \Pi$ and by choosing the optimal output model $\mathbf{P}(a_t|\underline{ao}_{<t})$ are not exactly the same. This is because choosing $\pi$ always leads to deterministic conditional probabilities for the output model, while directly choosing the output model allows choosing even non-deterministic output models. However, these optimal non-deterministic output models always arise as arbitrary mixtures of optimal deterministic output models, and thus are not "truly probabilistic". In particular, one can *always* choose an optimal deterministic output model. This will become clear in Section 3.2.3. □

### 3.2.2 I/O Model

In the previous section we have seen that purposeful behavior entails holding "implicit beliefs" about the input stream which are formalized by a set of conditional probabilities $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$. This conditional probability measure will turn out to be a fundamental part of the information-theoretic characterization of autonomous agents. The next definition extends the output model introduced in Chapter 2 with a model of the input stream.

**Definition 9 (Input Model)** An **input model** of an agent is a set of conditional probabilities

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) \quad \text{for all} \quad \underline{ao}_{\leq t} \in \mathcal{Z}^\diamond.$$

inducing a unique probability measure $\mathbf{P}$ over $\mathcal{O}^T$ conditioned on $\mathcal{A}^T$. □

**Remark 3** Note that an input model for an agent is formally equivalent to an output model of an environment. □

**Definition 10 (I/O Model)** An **I/O model** of an agent is an output model paired with an input model, i.e. a set of conditional probabilities

$$\mathbf{P}(a_t|\underline{ao}_{<t}) \quad \text{and} \quad \mathbf{P}(o_t|\underline{ao}_{<t}a_t) \quad \text{for all} \quad \underline{ao}_{<t} \in \mathcal{Z}^\diamond.$$

inducing a unique probability measure $\mathbf{P}$ over $\mathcal{Z}^T$ given by

$$\mathbf{P}(\underline{ao}_{<t}a_t) := \mathbf{P}(\underline{ao}_{<t})\mathbf{P}(a_t|\underline{ao}_{<t}) \quad \text{and} \quad \mathbf{P}(\underline{ao}_{\leq t}) := \prod_{\tau=1}^{t} \mathbf{P}(a_\tau|\underline{ao}_{<\tau})\mathbf{P}(o_\tau|\underline{ao}_{<\tau}a_\tau).$$

□

The intuitive interpretation of the I/O model is as follows. The $\mathbf{P}(a_t|\underline{ao}_{<t})$ are the conditional probabilities that the autonomous system follows to generate actions, i.e. they constitute **propensities**. It is common to call the output model the **policy** of the I/O model. In contrast, the $\mathbf{P}(o_t|\underline{ao}_{<t})$ are the conditional probabilities that the autonomous system uses to represent its assumptions, beliefs and/or predictions about the observations, i.e. they constitute **plausibilities**. It is common to call the input model the **predictor** of the I/O model.

**Definition 11 ("Knows")** An agent $\mathbf{P}$ (a I/O model) is said to **know** its environment $\mathbf{Q}$ (a output model) iff

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) = \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) \quad \text{for all} \quad \underline{ao}_{\leq t} \in \mathcal{Z}^\diamond.$$

□

In other words, an agent knows its environment if it perfectly predicts its behavior. Although this assumption is rarely met in practice, it is often a useful approximation. An agent that is constructed using the maximum SEU principle is said to be **optimal** (with respect to its input model).

### 3.2.3 Bellman Optimality Equations

The variational problem in (3.2) can be formulated as recursive system of equations collectively known as **Bellman optimality equations** (Bellman, 1957). Given a utility function $\mathbf{U}$ over $\mathcal{Z}^T$ and given a set of conditional probabilities $\mathbf{P}(o_t|\underline{ao}_{<t})$, the problem of finding the optimal output model can be graphically represented as a decision tree (see Figure 3.8).

There are two equivalent ways of conceptualizing this decision problem. The first is to find a policy such that, when executed, it will eventually lead to an optimal

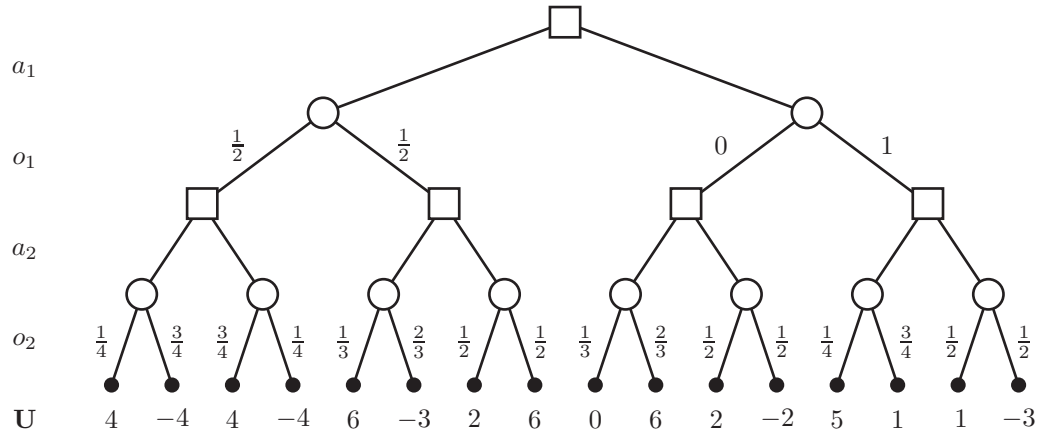



**Figure 3.8:** A Decision Tree. Square nodes (□) are action nodes, round nodes (◯) are observations nodes and leaves (•) are full I/O histories, which correspond to the consequences of the decision problem. The transition probabilities in observation nodes are given by the input model $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$, and the utilities of the consequences by the $\mathbf{U}(\underline{ao}_{\leq T})$. The problem consists in choosing the transition probabilities in action nodes (i.e. the $\mathbf{P}(a_t|\underline{ao}_{<t})$) such that the expected utility of the resulting I/O history is maximized.

"final state" having a maximum utility. The second is to find a policy such that, when executed, collects the maximum expected sum of "instantaneous rewards" before eventually reaching a final state.

**Definition 12 (Optimality Equations for Utilities)** Let $\mathbf{U}$ be a utility function over $\mathcal{Z}^T$ and let the $\mathbf{P}(o_t|\underline{ao}_{<t})$ be a set of conditional probabilities. Define the *future utility function* $F : \mathcal{Z}^\diamond \to \mathbb{R}$ recursively as

$$\begin{aligned} F(\underline{ao}_{\leq T}) &:= \mathbf{U}(\underline{ao}_{\leq T}) \\ F(\underline{ao}_{<t}) &:= \max_{a_t} \sum_{o_t} \mathbf{P}(o_t|\underline{ao}_{<t}a_t) F(\underline{ao}_{\leq t}), \end{aligned}$$

for all $\underline{ao}_{<t} \in \mathcal{Z}^\diamond$ with $t \leq T$. Then, the solution to the variational problem in (3.2) is the output model given by

$$\mathbf{P}(a_t|\underline{ao}_{<t}) := \delta_{a_t}^{a_t^*}, \quad \text{where} \quad a_t^* := \arg\max_{a_t} \sum_{o_t} \mathbf{P}(o_t|\underline{ao}_{<t}a_t) F(\underline{ao}_{\leq t}). \qquad \square$$

**Remark 4** If there are several maximizing actions, then one can choose the action that lexicographically precedes all the other ones for definiteness. □

It can be verified that the previous definition characterizes the optimal output model. The Bellman optimality equations of Definition 12 suggest a solution algorithm called Expectimax (Michie, 1966; Russell and Norvig, 2009) illustrated in Figure 3.9.

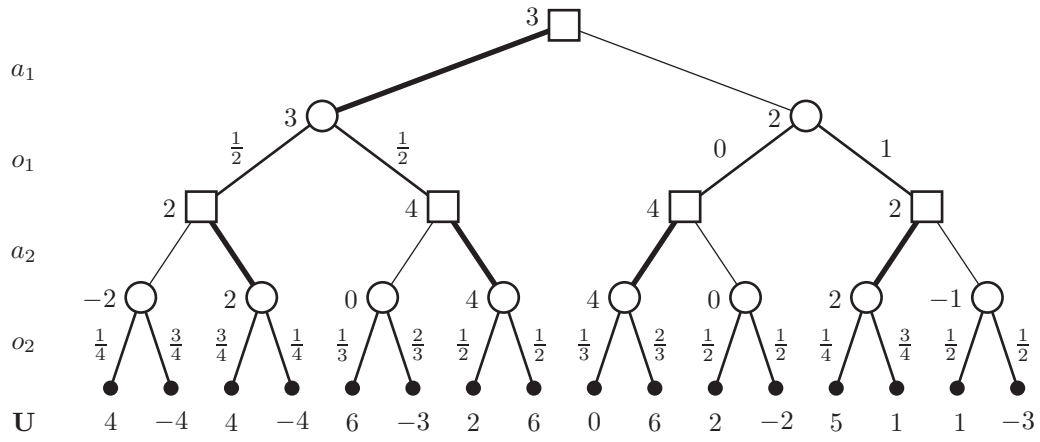



**Figure 3.9:** Solution of the Decision Tree from Figure 3.8. Starting from the leaves and working up until reaching the root, nodes are successively labeled with either the expected value of their children nodes (in observation nodes) or the maximum value of their children nodes (in action nodes). The transitions chosen in the action nodes form the solution. Note that the values written in action nodes and in the leaves correspond to the future utilities (Definition 12), and that the differences of two subsequent future utilities correspond to the rewards (Definition 14).

As anticipated previously, one can formulate this recursion in terms of collecting "instantaneous rewards" in each interaction. This formulation is useful for modeling situations where there is a feedback signal communicating the autonomous system how well it is performing. We introduce the definition of rewards.

**Definition 13 (Rewards)** Let $F : \mathcal{Z}^\diamond \to \mathbb{R}$ the future utility from Definition 12. Then, define the **reward function** $r$ as

$$r(\underline{ao}_t|\underline{ao}_{<t}) := F(\underline{ao}_{\leq t}) - F(\underline{ao}_{<t})$$

for all $\underline{ao}_{\leq t} \in \mathcal{Z}^\diamond$ where $t \geq 1$. □

In a reward-based decision problem, the goal is to maximize the cumulative sum over rewards, i.e. the quantity

$$\sum_{t=1}^{T} r(\underline{ao}_t|\underline{ao}_{<t}) = F(\underline{ao}_{\leq T}) - F(\epsilon).$$

Although rewards are mathematically formalized as differences in future utility, decision problems based on rewards are most naturally stated by directly specifying the reward function, i.e. without explicitly deriving it from an overall utility function. The associated optimality equations follow.

**Definition 14 (Optimality Equations for Rewards)** Let $r$ be a reward function and let the $\mathbf{P}(o_t|\underline{ao}_{<t})$ be a set of conditional probabilities. Define the **value function** $V : \mathcal{Z}^\diamond \to \mathbb{R}$ recursively as

$$V(\underline{ao}_{\leq T}) := 0$$
$$V(\underline{ao}_{<t}) := \max_{a_t} \sum_{o_t} \mathbf{P}(o_t|\underline{ao}_{<t}a_t)\big(r(\underline{ao}_t|\underline{ao}_{<t}) + V(\underline{ao}_{\leq t})\big),$$

for all $\underline{ao}_{<t} \in \mathcal{Z}^\diamond$ with $t < T$. Then, the solution is the output model given by

$$\mathbf{P}(a_t|\underline{ao}_{<t}) := \delta_{a_t}^{a_t^*}, \quad \text{where} \quad a_t^* := \arg\max_{a_t} \sum_{o_t} \mathbf{P}(o_t|\underline{ao}_{<t}a_t)\big(r(\underline{ao}_t|\underline{ao}_{<t}) + V(\underline{ao}_{\leq t})\big). \square$$

**Remark 5** Note that the choice of the policy $\mathbf{P}(a_t|\underline{ao}_{<t})$ *causally precedes* the realization of the stochastic process $a_1, o_1, a_2, o_2, \ldots, a_T, o_T$ in all the formulations we have seen so far. This means that the agent has to "know how it will act under all potential situations" before having interacted with the environment even once. We will relax this assumption in Chapter 9. $\square$

### 3.2.4 Subjective versus True Expected Utility

In the previous section we have seen how to solve a decision problem. This design method makes especially sense when the agent knows its environment, because

$$\begin{aligned}\sum_{\underline{ao}_{\leq T}} \mathbf{G}(\underline{ao}_{\leq T})\mathbf{U}(\underline{ao}_{\leq T}) &= \sum_{\underline{ao}_{\leq T}} \left(\prod_{t=1}^{T} \mathbf{P}(a_t|\underline{ao}_{<t})\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)\right)\mathbf{U}(\underline{ao}_{\leq T}) \\ &= \sum_{\underline{ao}_{\leq T}} \left(\prod_{t=1}^{T} \mathbf{P}(a_t|\underline{ao}_{<t})\mathbf{P}(o_t|\underline{ao}_{<t}a_t)\right)\mathbf{U}(\underline{ao}_{\leq T}) \\ &= \sum_{\underline{ao}_{\leq T}} \mathbf{P}(\underline{ao}_{\leq T})\mathbf{U}(\underline{ao}_{\leq T}),\end{aligned}$$

where $\mathbf{G}$ is the generative probability measure (Definition 3), and hence the true expected utility arising from the interaction between the agent $\mathbf{P}$ and the environment $\mathbf{Q}$ coincides with the subjective expected utility of the agent $\mathbf{P}$. As we have pointed out in the context of Definition 11, this is a requirement that is hardly fulfilled in practice. Finding the optimal policy given that the agent knows the environment constitutes the paradigm of **optimal control**. Optimal control is an important problem class having application in many real-life situations.

Obviously, if the agent does not know its environment, then the true expected utility and the subjective expected utility do not coincide and in fact little can be said about how these two quantities compare, i.e.

$$\sum_{\underline{ao}_{\leq T}} \mathbf{G}(\underline{ao}_{\leq T})\mathbf{U}(\underline{ao}_{\leq T}) \overset{?}{\lessgtr} \sum_{\underline{ao}_{\leq T}} \mathbf{P}(\underline{ao}_{\leq T})\mathbf{U}(\underline{ao}_{\leq T}).$$

Thus, it seems that rational agents are somewhat fragile, in the sense that their performance seems to depend strongly on knowing the environment they are facing. This begs the question of whether we can design more robust agents, e.g. agents that perform well even if they don't know what environment they are facing. This is the question addressed in the next chapter.

## 3.3 Historical Remarks & References

The concept of utility has become an integral part of the standard vocabulary in economics, control theory and artificial intelligence. However, its development has been slow and controversial. The roots of decision theory can be traced back as far as before the 18th century. Gamblers were thought to evaluate random ventures based on their expected returns (i.e. expected monetary payoffs). This led to the famous *St. Petersburg paradox*: a gamble with infinite expected return, but nevertheless intuitively considered to be worth only a very small amount of money. Bernoulli (1738) presented a solution to this paradox by introducing the notion of utilities: an outcome is not worth its nominal (i.e. monetary) value, but a subjective transformation of this value. Accordingly, the valuation of a gamble would then reduce to the calculation of its expected utility, which decomposed this valuation into a sum of utilities of outcomes weighted by their likelihoods. Surprisingly though, this idea did generally not pick up with early economists. It was only after Knight (1921) made a case suggesting that risk, uncertainty and utility might be relevant for economic analysis that the leading economists finally took these concepts seriously into account.

Ramsey (1926) suggested a way of deriving a consistent theory of choice under uncertainty that could isolate subjective probabilities (i.e. beliefs) from preferences. This idea got later independently developed by De Finetti (1937). In his formulation, he proposed a thought experiment where a decision maker is required to set the fee for a gamble involving the likelihood of an event. He then showed that the decision maker has to choose a fee that is in accordance with his subjective belief in the likelihood of the event in order to avoid being a victim of a *Dutch book*, i.e. a strategy that an opponent can use in order to systematically win against him.

The first mathematical formalization of the expected utility hypothesis came with von Neumann and Morgenstern (1944). In their work, a decision maker is formalized as having preferences over lotteries (i.e. probability distributions over outcomes). von Neumann and Morgenstern have shown that if this preference relation follows certain consistency axioms, then there exists an "essentially" unique utility function over lotteries. Utilities over outcomes are then derived from utilities of deterministic lotteries. Note that this formalization is based on at least two important assumptions. First, probabilities are assumed to be "objective", i.e. they are given by Nature and cannot be influenced by the decision maker. Second, lotteries logically precede outcomes, which seems to suggest that decision makers really get utility from distributions, not from outcomes.

The ideas of subjective utility and probability culminated in the theory of subjective expected utility developed by Savage (1954). Savage combined de Finetti's and von Neumann-Morgenstern's ideas into the theory of decision making under uncertainty presented in this chapter. This theory is by many considered the biggest achievement in rational decision making, being the standard mathematical framework until today. Later, Anscombe and Aumann (1963) provided a simpler derivation by extending von Neumann and Morgenstern's approach with subjective probabilities. The exposition in Section 3.1 follows closely the one presented in Kreps (1988, Chapter 9). A mathematically rigorous treatment is given in Fishburn (1982).

In control theory, Bellman (1957) discovered the optimality equations named after him.

These equations constitute a necessary condition for optimality. As it has been explained in Section 3.2.3, the advantage of this condition is that it expresses the solution of a large optimization problem as a recursive combination of smaller optimization problems.

Bellman's work turned out to be very influential, and his method quickly established itself as the standard in the control community. In particular, Bellman optimality equations have been mainly applied to efficiently solve **Markov decision problems** (MDPs). An **MDP** is a probabilistic model of a special class of environments, and it emphasizes the generative aspect of observations. From a computational point of view, a finite MDP is a finite state machine with stochastic transitions influenced by the actions of the agent, and where the computational states correspond exactly to the observations that the agent can make. The latter assumption allows devising an efficient computational algorithm, called **dynamic programming**, to find the optimal policy in (linear) running time $\mathrm{O}(|\mathcal{Z}|\,T)$ (Tsitsiklis, 2007). Because of the wide range of applicability of MDPs and the low computational cost of finding their solution, MDPs have become almost synonymous with control problems. Later, as researchers have attempted to tackle more realistic problems, they increasingly realized the shortcomings of the MDP assumptions. As a response to this, the MDP model has been generalized to partially observable MDPs (POMDPs), where observations are allowed to be functions of the computational state (Russell and Norvig, 2009). However, the computational complexity of solving a POMDP grows exponentially with the number of computational states. It is worth point out that the POMDP formulation and the formulation presented in this thesis are essentially equivalent, with the difference that POMDPs emphasize a particular generative model for observations, while the generative mechanism is left open in this thesis.

Finally, while SEU theory makes a strong case for being a normative[1] model of decision making, it is not flawless. In fact, there is strong experimental evidence that disconfirms the validity of SEU as a normative model in systematic ways—most notably *Ellberg's Paradox*, which has puzzled even Savage himself (Ellsberg, 1961). The interested reader can refer to the specialized literature (Machina, 1987).

[1]Normative in the sense that a rational decision maker would revise his decisions if they do not follow the maximum SEU principle.

# Chapter 4

# Learning

In the previous chapter, we have seen how to solve an optimal control problem, i.e. how to construct the policy of an agent when the environment is known. Examples of this situation include hitting a target with a cannon under known weather conditions, solving a maze having its map and controlling a robotic arm in a manufacturing plant. However, when the environment is unknown, then the agent needs an adaptive policy. For example, shooting the cannon lacking the appropriate measurement equipment, finding the way out of an unknown maze and designing an autonomous robot for Martian exploration. In all these examples, the agent needs to "learn" its environment on-the-fly in order to optimize its performance. How to design adaptive agents within SEU theory is the subject matter of this chapter.

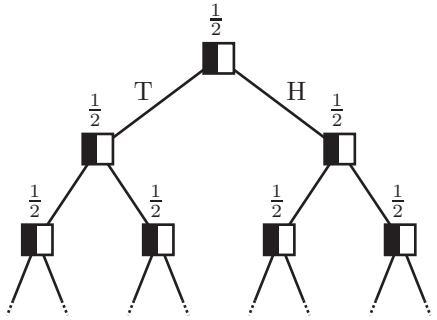



**Figure 4.1:** A Fixed Predictor.

To illustrate learning, consider the following example. The task is to predict the outcomes of a sequence of coin tosses. The coin has an unknown bias drawn from $\theta \in [0, 1]$. Observations $o_t$ are drawn from $\mathcal{O} := \{\text{H}, \text{T}\}$, and beliefs are $\mathbf{P}(o_t|o_{<t})$ for any $o_{\leq t} \in \mathcal{O}^*$ (actions are omitted in this case). Consider a first belief given by

$$\mathbf{P}(o_t = \text{H}|o_{<t}) = \mathbf{P}(o_t = \text{T}|o_{<t}) = \frac{1}{2}.$$

This is illustrated in Figure 4.1. This predictor does not learn anything, because it maintains a fixed prediction over the next outcome, despite the past observations. In contrast, consider the belief illustrated in Figure 4.2. This predictor learns from experience, because it adjusts its prediction based on past observations. Two important properties should be emphasized. First, it uses the past observations to find a suitable prediction for the next coin toss (that is, within a predefined class of predictors, namely, the Bernoulli processes[1]). Second, the predictor "accumulates evidence", since the "amount of change" is decreasing.

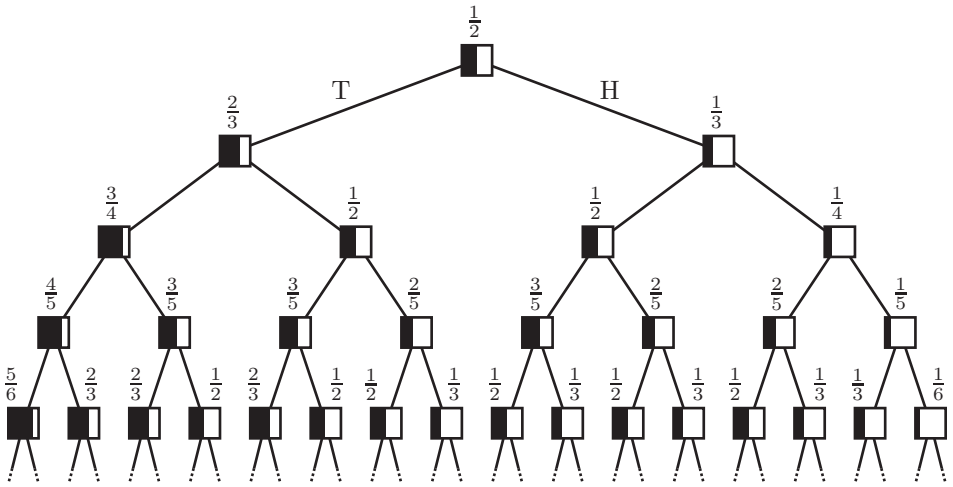



**Figure 4.2:** An Adaptive Predictor.

Let $h$ denote the number of times a head has been observed at time $t$. The probabilities in the second predictor have been chosen as

$$\mathbf{P}(o_{t+1} = \mathrm{H} | o_{\leq t}) = \frac{h+1}{t+2},$$

known as the *rule of succession*. It is easy to see that this rule converges to the right odds with enough coin tosses. The choice of this predictor is not arbitrary: it can be given a statistical justification that will be developed in this chapter.

## 4.1 Bayesian Probability Theory

So far, we have given probabilities two usages: to specify a generative law (i.e. propensities) and to specify degrees of belief (i.e. plausibilities). While both usages obey the axioms of probability, it would be desirable to have an explanatory framework for plausibilities that is intuitively more appealing. Such a framework is Bayesian probability

[1] An i.i.d. sequence of binary random variables with a fixed bias.

theory. Simply put, Bayesian probability theory is a framework that extends logic for reasoning under uncertainty. The presentation of Bayesian probability theory outlined in the following is an original contribution of this thesis.

### 4.1.1 Reasoning under Certainty

Logic is the most important framework of reasoning (under certainty). Here, it is rephrased in set-theoretic terms[2]. As will be seen, this facilitates its extension to a framework for reasoning under uncertainty.

Let $\Omega$ be a set of **outcomes**, which is assumed to be finite for simplicity. A subset $A \subset \Omega$ is an **event**. Let $^c$, $\cup$ and $\cap$ be the set-operations of **complement**, **union** and **intersection** respectively. Let $\mathcal{F}$ be an **algebra**, i.e. a set of events obeying the axioms

A1. $\mathcal{F} \neq \varnothing$.

A2. $A \in \mathcal{F} \quad \Rightarrow \quad A^c \in \mathcal{F}$.

A3. $A, B \in \mathcal{F} \quad \Rightarrow \quad A \cup B \in \mathcal{F}$.

In this framework, an outcome $\omega \in \Omega$ represents a state of affairs and an event $A \in \mathcal{F}$ a proposition. For instance, if two coins are tossed, then there are four outcomes $\Omega = \{\text{HH, HT, TH, TT}\}$. A possible outcome is $w = \text{HT}$ (i.e. first head, then tail), and an event is $A = \{\text{HH, HT, TH}\}$ (i.e. there is a head). Hence, a singleton $\{\omega\} \in \mathcal{F}$ is an irreducible (i.e. atomic) proposition about the world. The set-operations $^c$, $\cup$ and $\cap$ correspond to the logical connectives of $\neg$ (negation), $\vee$ (disjunction) and $\wedge$ (conjunction) respectively. They allow the construction of complex propositions from simpler ones. An algebra is a system of propositions that is closed under negation and disjunction (and hence is closed under conjunction as well), i.e. it comprises all propositions that the reasoner might entertain.

**Remark 6** A consequence of the axioms is that both the universal event $\Omega$ and the impossible event $\varnothing$ are in $\mathcal{F}$. □

The objective of logic is to allow the reasoner to conclude the veracity of events given information. Let $\mathcal{V} := \{1, 0, ?\}$ be the set of **truth states**, where 1 is **true**, 0 is **false**, and ? is **uncertain** (but known to be either true or false). From these, $\{1, 0\}$ are called **truth values**. The **truth function** is the set function $\mathbf{T}$ over $\mathcal{F} \times \mathcal{F}$ defined as

$$A, B \in \mathcal{F}, \quad \mathbf{T}(A|B) = \begin{cases} 1 & \text{if } B \subset A, \\ 0 & \text{if } A \cap B = \varnothing, \\ ? & \text{else.} \end{cases}$$

[2]Strictly speaking, this set-theoretic logic is "a logic within logic", since set theory is based on standard logic.

Furthermore, define the shorthand $\mathbf{T}(A) := \mathbf{T}(A|\Omega)$. The quantity $\mathbf{T}(A|B)$ stands for the "truth value of event $A$ given that event $B$ is true". Accordingly, the knowledge of the reasoner about the facts of the world is represented by his truth function and his algebra. From his point of view, a proposition can be either true, false or uncertain (i.e. having an unresolved truth value given his knowledge). Understanding the definition of the truth function is simple. Claiming that an event $B \in \mathcal{F}$ *is true* means that one of its members $\omega \in B$ is the current outcome/state of affairs. Thus, for instance, claiming that the statement "there is some head" is true means that $\omega \in \{$HH, HT, TH$\}$. Hence the veracity of $A$ given $B$ is evaluated as follows (Figure 4.3): if $A$ contains every outcome in $B$ then it must be true as well; if $A$ is known not to contain any of $B$'s outcome then it must be false; and if $A$ contains only part of $B$ then it cannot be resolved, since knowing that $\omega \in B$ does neither imply that $\omega \in A$ nor $\omega \in A^c$. The definition of a truth space follows.

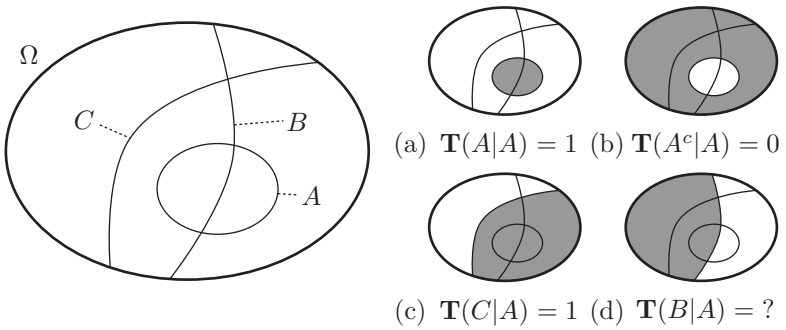



**Figure 4.3:** A truth space. It is known that the true outcome $\omega \in \Omega$ is in $A$. Hence, (a) the event $A$ is true and (b) its complement $A^c$ is false. (c) Any event that contains a true event is true as well. (d) An event that contains only part of a true event is uncertain.

**Definition 15 (Truth Space)** A **truth space** is a tuple $(\Omega, \mathcal{F}, \mathbf{T})$ where: $\Omega$ is a set of outcomes, $\mathcal{F}$ is an algebra over $\Omega$ and $\mathbf{T} : \mathcal{F} \times \mathcal{F} \to \mathcal{V}$ is a truth function. □

The intuitive meaning of a truth space is as follows. Nature arbitrarily selects an outcome $\omega \in \Omega$. (This choice is *not* governed by a generative law.) *Subsequently*, the reasoner performs a measurement: he chooses a set $B$ and nature reveals to him whether $\omega \in B$ or not. Accordingly, the reasoners infers the veracity of any event $A \in \mathcal{F}$ by evaluating either $\mathbf{T}(A|B)$ (if $\omega \in B$) or $\mathbf{T}(A|B^c)$ (if $\omega \notin B$).

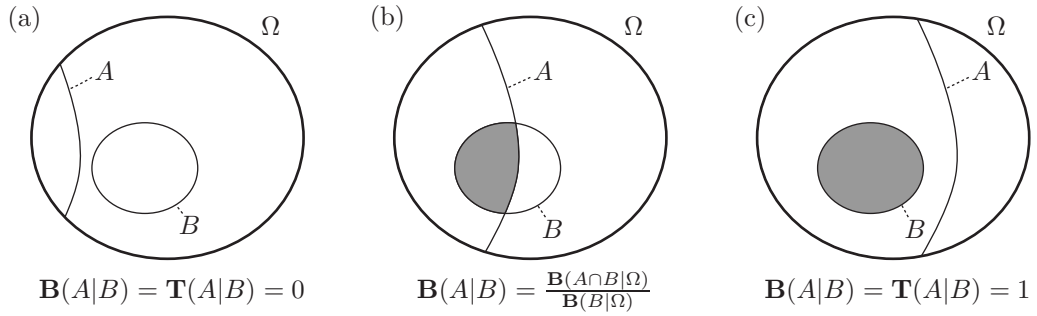
Several measurements are combined as a conjunction. Thus, if the reasoner learns that $\omega$ is in $B_1$, $B_2$, ..., and $B_t$ after performing $t$ measurements, then the truth value is $\mathbf{T}(A|B_1 \cap \cdots \cap B_t)$ for any $A \in \mathcal{F}$.

**Remark 7** Knowing that $\omega \in \Omega$ does not resolve uncertainty, i.e. $\mathbf{T}(A|\Omega) = ?$ for any $A \in \mathcal{F} \setminus \{\Omega, \varnothing\}$, while knowing that $\omega \in \{\omega\}$ resolves all uncertainty, i.e. $\mathbf{T}(A|\{\omega\}) \in \{0, 1\}$ for any $A \in \mathcal{F}$. □

**Remark 8** The set relation $B \subset A$ corresponds to the logical relation $B \Rightarrow A$. Since an algebra is an encoding of how sets are contained within each other, it should be clear that an algebra is essentially a system of implications. □

### 4.1.2 Reasoning under Uncertainty

Unlike logic, Bayesian probability theory allows reasoning under uncertainty. For this end, it provides a consistent mechanism to replace the uncertainty state ? with a numerical value in the interval $[0, 1]$ representing degrees of truth, belief or plausibility.




$\mathbf{B}(A|B) = \mathbf{T}(A|B) = 0$ $\quad$ $\mathbf{B}(A|B) = \frac{\mathbf{B}(A \cap B|\Omega)}{\mathbf{B}(B|\Omega)}$ $\quad$ $\mathbf{B}(A|B) = \mathbf{T}(A|B) = 1$

**Figure 4.4:** Extension of Truth Function.

The goal is to find a suitable definition of a quantity $\mathbf{B}(A|B)$ meaning "the degree of belief in event $A$ given that event $B$ is true" that is consistent with the truth function when it is certain, i.e. $\mathbf{B}(A|B) := \mathbf{T}(A|B)$ if $\mathbf{T}(A|B) \in \{0, 1\}$. Consider the three situations in Figure 4.4. (a) In the case $A \cap B = \varnothing$, we impose $\mathbf{B}(A|B) := \mathbf{T}(A|B) = 0$. (c) In the case $B \subset A$, we impose $\mathbf{B}(A|B) := \mathbf{T}(A|B) = 1$. (b) In the intermediate case where $\mathbf{T}(A|B) = ?$, the event $A$ only partially covers the members of $B$. If one interprets the quantity $\mathbf{B}(C|D)$ as "the fraction of $D$ contained in $C$", then one can characterize $\mathbf{B}(A|B)$ with the relation

$$\mathbf{B}(A|B) = \frac{\mathbf{B}(A \cap B|\Omega)}{\mathbf{B}(B|\Omega)}$$

as long as $\mathbf{B}(B|\Omega) > 0$. It is easy to see that this formula generalizes correctly to the border cases, since $\mathbf{B}(A|B) = \frac{0}{\mathbf{B}(B|\Omega)} = 0$ when $A \cap B = \varnothing$ and $\mathbf{B}(A|B) = \frac{\mathbf{B}(B|\Omega)}{\mathbf{B}(B|\Omega)} = 1$ when $B \subset A$. Noting that $B = B \cap \Omega$ and rearranging terms, one gets

$$\mathbf{B}(A \cap B|\Omega) = \mathbf{B}(B|\Omega)\, \mathbf{B}(A|B \cap \Omega).$$

We demand this relation to hold under any restriction to a "universal" set $C \in \mathcal{F}$, not only when it is restricted to $\Omega$. Thus, replacing $\Omega$ by $C$ one obtains

$$\mathbf{B}(A \cap B|C) = \mathbf{B}(B|C)\, \mathbf{B}(A|B \cap C),$$

which is known as the **product rule** for beliefs. Following a similar reasoning, we impose that for any event $A \in \mathcal{F}$, the sum of the degree of belief in $A$ and its complement $A^c$ must be true under any condition $B$, i.e.

$$\mathbf{B}(A|B) + \mathbf{B}(A^c|B) = 1,$$

which is known as the **sum rule** for beliefs. In summary, we impose the following axioms for beliefs (Jaynes and Bretthorst, 2003).

**Definition 16 (Belief axioms)** Let $\Omega$ be a set of outcomes and let $\mathcal{F}$ be an algebra over $\Omega$. A set function $\mathbf{P}$ over $\mathcal{F} \times \mathcal{F}$ is a **belief function** iff

B1. $A, B \in \mathcal{F}, \quad \mathbf{B}(A|B) \in [0, 1].$

B2. $A, B \in \mathcal{F}, \quad \mathbf{B}(A|B) = 1 \quad$ if $B \subset A$.

B3. $A, B \in \mathcal{F}, \quad \mathbf{B}(A|B) = 0 \quad$ if $A \cap B = \varnothing$.

B4. $A, B \in \mathcal{F}, \quad \mathbf{B}(A|B) + \mathbf{B}(A^c|B) = 1.$

B5. $A, B, C \in \mathcal{F}, \quad \mathbf{B}(A \cap B|C) = \mathbf{B}(A|C)\,\mathbf{B}(B|A \cap C).$ □

Furthermore, define the shorthand $\mathbf{B}(A) := \mathbf{B}(A|\Omega)$. Axiom B1 states that degrees of belief are real values in the unit interval $[0, 1]$. Axioms B2 and B3 equate the belief and the truth function under certainty. Axioms B4 and B5 are the structural requirements under uncertainty discussed above. Accordingly, one defines a belief space as follows.

**Definition 17 (Belief Space)** A **belief space** is a tuple $(\Omega, \mathcal{F}, \mathbf{B})$ where: $\Omega$ is a set of outcomes, $\mathcal{F}$ is an algebra over $\Omega$ and $\mathbf{B} : \mathcal{F} \times \mathcal{F} \to [0, 1]$ is a belief function. □

The intuitive meaning of a belief space is analogous to a truth space. Nature arbitrarily selects an outcome $\omega \in \Omega$. *Subsequently*, the reasoner performs a measurement: he chooses a set $B$ and nature reveals to him whether $\omega \in B$ or not. Accordingly, the reasoner infers the degree of belief in any event $A \in \mathcal{F}$ by evaluating either $\mathbf{B}(A|B)$ (if $\omega \in B$) or $\mathbf{B}(A|B^c)$ (if $\omega \notin B$).

**Remark 9** The word "subsequently", that has been emphasized for the second time now, is crucial. When the reasoner performs his measurements, the outcome is already determined. In Chapter 8, we will relax this assumption by allowing outcomes that are only partially determined or jointly determined by the reasoner him-/herself. □

**Remark 10** Note that logical truth is *not* the same as probabilistic truth. That is

$$\begin{aligned} \mathbf{T}(A|B) = 1 \quad &\Rightarrow \quad \mathbf{B}(A|B) = 1, \\ \text{but} \quad \mathbf{B}(A|B) = 1 \quad &\Rightarrow \quad \mathbf{T}(A|B) \in \{?, 1\}. \end{aligned}$$

In particular, the case $\mathbf{B}(A|B) = 1$ and $\mathbf{T}(A|B) = ?$ occurs when $\mathbf{B}$ is chosen such that the set $B \setminus A$ is non-empty but has no probability mass relative to $B$. In other words, if $\mathbf{B}(B \setminus A|B) := 0$, then $\mathbf{B}(A \cap B|B) = \mathbf{B}(A|B) = 1$, even though $B \not\subset A$. □

*An easy but fundamental result is that the axioms of belief are equivalent to the axioms of probability*[3] *(Jaynes and Bretthorst, 2003).* This simple observation is what constitutes the foundation of Bayesian probability theory.

### 4.1.3 Bayes' Rule

We now return to the central topic of this chapter. Suppose the reasoner has uncertainty over a set of competing hypotheses about the world. Subsequently, he makes an observation. He can use this observation to update his beliefs about the hypotheses. The following theorem explains how to carry out this update.

**Theorem 2 (Bayes' Rule)** *Let* $(\Omega, \mathcal{F}, \mathbf{B})$ *be a belief space. Let* $\{H_1, \ldots, H_N\}$ *be a partition of* $\Omega$*, and let* $D \in \mathcal{F}$ *be an event such that* $\mathbf{B}(D) > 0$*. Then, for all* $n \in \{1, \ldots, N\}$*,*

$$\mathbf{B}(H_n|D) = \frac{\mathbf{B}(D|H_n)\,\mathbf{B}(H_n)}{\mathbf{B}(D)} = \frac{\mathbf{B}(D|H_n)\,\mathbf{B}(H_n)}{\sum_m \mathbf{B}(D|H_m)\,\mathbf{B}(H_m)}.$$

*This is known as* ***Bayes' rule****.* □

The interpretation is as follows. The $H_1, \ldots, H_N$ represent $N$ mutually exclusive **hypotheses**, and the event $D$ represents a new observation or **data**. Initially, the reasoner holds a **prior belief** $\mathbf{B}(H_n)$ over each hypothesis $H_n$. Subsequently, he incorporates the observation of the event $D$ and arrives at a **posterior belief** $\mathbf{B}(H_n|D)$ over each hypothesis $H_n$. Bayes' rule states that this update can be seen as combining the prior belief $\mathbf{B}(H_n)$ with the **likelihood** $\mathbf{B}(D|H_n)$ of observation $D$ under hypothesis $H_n$. The denominator $\sum_m \mathbf{B}(D|H_m)\mathbf{B}(H_m) = \mathbf{B}(D)$ just plays the role of a normalizing constant (Figure 4.5).

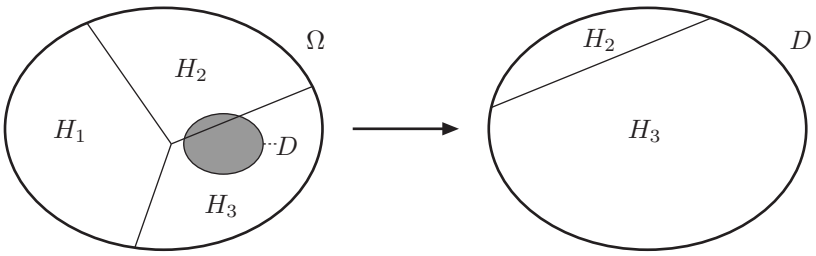



**Figure 4.5:** Schematic Representation of Bayes' Rule. The prior belief in hypotheses $H_1$, $H_2$ and $H_3$ is roughly uniform. After conditioning on the observation $D$, the belief in hypothesis $H_3$ increases significantly.

[3]More precisely, the axioms of beliefs as stated here imply the axioms of probability for finitely additive measures over finite algebras. Furthermore, the axioms of beliefs also specify a unique version of the conditional probability measure.

Bayes' rule naturally applies to a sequential setting. Incorporating a new observation $D_t$ after having observed $D_1, D_2, \ldots, D_{t-1}$ updates the beliefs as

$$\mathbf{B}_{t+1}(H_n) := \mathbf{B}(H_n|D_1 \cap \cdots \cap D_t) = \frac{\mathbf{B}_t(D_t|H_n)\,\mathbf{B}_t(H_n)}{\sum_m \mathbf{B}_t(D_t|H_m)\,\mathbf{B}(H_m)},$$

where for the $t$-th update,

$$\mathbf{B}_t(H_n) := \mathbf{B}(H_n|D_1 \cap \cdots \cap D_{t-1}) \quad \text{and} \quad \mathbf{B}_t(D_t|H_n) := \mathbf{B}(D_t|H_n \cap D_1 \cap \cdots \cap D_{t-1})$$

play the role of the prior belief and the likelihood respectively. Note that

$$\mathbf{B}(D_1 \cap \cdots \cap D_t|H_n) = \prod_{\tau=1}^{t} \mathbf{B}(D_\tau|H_n \cap D_1 \cap \cdots \cap D_{\tau-1}),$$

and hence each hypothesis $H_n$ naturally determines a probability measure $\mathbf{B}(\cdot|H_n)$ over sequences of observations.

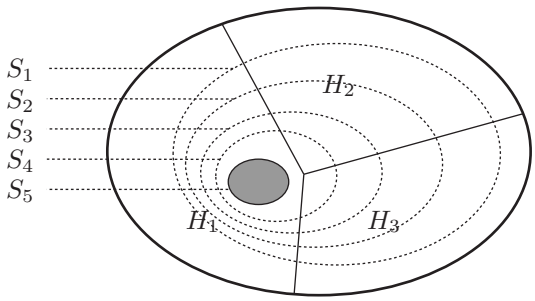



**Figure 4.6:** Progressive refinement of the accuracy of the joint observation. The sequence of observations $D_1, \ldots, D_5$ leads to refinements $S_1, S_2, \ldots, S_5$, where $S_t = D_1 \cap \cdots \cap D_t$. Note that $S_5 \subset H_1$ and therefore $\mathbf{B}(H_1|S_5) = 1$, while $\mathbf{B}(H_2|S_5) = \mathbf{B}(H_3|S_5) = 0$.

A smaller event $D$ corresponds to a more "accurate" observation. Hence, making a new observation $D'$ necessarily improves the accuracy, since

$$D \supset D \cap D'.$$

In some cases, the accuracy of an observation (or sequence of observations) can be so high that it uniquely identifies a hypothesis (Figure 4.6).

The way Bayes' rule operates can be illustrated as follows. Consider a partition $\{X_1, \ldots, X_K\}$ of $\Omega$ and let $H_* \in \{H_1, \ldots, H_N\}$ be the true hypothesis, i.e. the outcome $\omega \in \Omega$ is drawn obeying propensities described by $\mathbf{B}(\cdot|H_*)$. The $X_k$ represent different observations the reasoner can make. If $\omega$ is drawn by Nature and reported to be in

$X_k$ (without revealing $\omega$ itself), then the log-posterior probability of hypothesis $H_n$ is given by

$$\log \mathbf{B}(H_n|X_k) = \underbrace{\log \mathbf{B}(X_k|H_n)}_{l_n} + \underbrace{\log \mathbf{B}(H_n)}_{p_n} - \underbrace{\log \mathbf{B}(X_k)}_{c}.$$

This decomposition highlights all the relevant terms for understanding Bayesian learning. The term $l_n$ is the log-likelihood of the data $X_k$. The term $p_n$ is the log-prior of hypothesis $H_n$, which is a way of representing the relative confidence in hypothesis $H_n$ prior to seeing the data. In practice, it can also be interpreted as (a) a complexity term, (b) the log-posterior resulting from "previous" inference steps, or (c) an initialization term for the inference procedure. The term $c$ is the log-probability of the data, which is constant over the hypotheses, and thus does not affect our analysis. Hence, log-posteriors are compared by their differences in $l_n + p_n$. Ideally, the log-posterior should be maximum for the true hypothesis $H_n = H_*$. However, since $\omega$ is chosen randomly, the observation $X_k$ is random, and hence the log-posterior $\log \mathbf{B}(H_n|X_k)$ is a random quantity too. If the variance of the log-posterior is high enough, then a particular realization of the data can lead to a log-posterior favoring some "wrong" hypotheses over the true hypothesis, i.e. $l_n + p_n > l_* + p_*$ for some $H_n \neq H_*$. In general, this is an unavoidable problem (that necessarily haunts *every* statistical inference method). Further insight can be gained by analyzing the expected log-posterior:

$$\underbrace{\sum_{X_k} \mathbf{B}(X_k|H_*) \log \mathbf{B}(X_k|H_n)}_{L_n} + \underbrace{\log \mathbf{B}(H_n)}_{P_n = p_n} - \underbrace{\sum_{X_k} \mathbf{B}(X_k|H_*) \log \mathbf{B}(X_k)}_{C}.$$

This reveals[4] that, *on average*, the log-likelihood $L_n$ is indeed maximized by $H_n = H_*$. Hence, the posterior belief will, on average, concentrate its mass on the hypotheses having high $L_n + P_n$.

## 4.2 Adaptive Optimal Control

The previous section introduced the conceptual framework of Bayesian probability theory to model reasoning under uncertainty. The aim of this section is to apply this framework to model adaptive autonomous systems. We first introduce a model to represent the uncertainty an agent has over its environment. Then, we show that this model of uncertainty also allows deriving a predictor over the input stream, which will be used as the input model of the agent. Finally, we show a convergence result for this predictive input model.

### 4.2.1 Bayesian Input Model

One can exploit the Bayesian interpretation of probabilities to construct adaptive autonomous systems. The Bayesian framework can be used to specify a belief model of

[4] For $p_i, q_i$ probabilities, $\sum_i p_i \log q_i$ is maximum when $q_i = p_i$ for fixed $p_i$.

agents having uncertainty about their environments. From this belief model, one can derive an adaptive predictor for the I/O model introduced in Chapter 3.

**Definition 18 (Bayesian Input Model)** Let $\Theta$ be a finite set and let $\mathcal{Z}^T$ be the set of interaction strings. A **Bayesian input model** of an agent is a set of conditional probabilities

$$P(\theta) \quad \text{and} \quad P(o_t|\theta, \underline{ao}_{<t}a_t) \quad \text{for all } \theta \in \Theta \text{ and all } \underline{ao}_{<t} \in \mathcal{Z}^\diamond,$$

uniquely specifying a conditional probability measure $P$ over $\Theta \times \mathcal{O}^T$ conditioned on $\mathcal{A}^T$ given by

$$P(\theta, o_{\leq t}|a_{\leq t}) = P(\theta) \prod_{\tau=1}^{t} P(o_\tau|\theta, \underline{ao}_{<\tau}a_\tau).$$

□

The Bayesian input model is a concise way of bundling together many hypotheses about the input stream. In fact, for each choice of $\theta$, the conditional probabilities $P(o_t|\theta, \underline{ao}_{<t}a_t)$ define an input model. One can derive a distribution over $\mathcal{O}^T$ conditioned on $\mathcal{A}^T$ by marginalizing over $\Theta$:

$$P(o_{\leq T}|a_{\leq T}) = \sum_\theta P(\theta) P(o_{\leq T}|\theta, a_{\leq T}). \tag{4.1}$$

Hence, the Bayesian input model defines a **mixture distribution** over the different input models, where the mixture weights are given by the $P(\theta)$.

### 4.2.2 Predictive Distribution

The Bayesian input model allows deriving an adaptive predictor, i.e. a predictor over the next observation $o_t$ that learns from the past observed I/O string $\underline{ao}_{<t}a_t$.

**Proposition 1 (Predictive Distribution)** *The probability of $o_t$ conditioned on the past $\underline{ao}_{<t}a_t$ is given by*

$$P(o_t|\underline{ao}_{<t}a_t) = \sum_\theta P(\theta|\underline{ao}_{<t})\, P(o_t|\theta, \underline{ao}_{<t}a_t), \tag{4.2}$$

*where the mixture weights $P(\theta|\underline{ao}_{\leq t})$ are given by the recursion*

$$P(\theta|\underline{ao}_{\leq t}) = \frac{P(o_t|\theta, \underline{ao}_{<t})\, P(\theta|\underline{ao}_{<t})}{\sum_{\theta'} P(o_t|\theta', \underline{ao}_{<t})\, P(\theta'|\underline{ao}_{<t})}.$$

*Equation (4.2) is known as the **predictive distribution**.* □

Proof First, $P(o_t|\underline{ao}_{<t}a_t)$ can be rewritten as

$$P(o_t|\underline{ao}_{<t}a_t) = \sum_\theta P(\theta|\underline{ao}_{<t}a_t)\, P(o_t|\theta, \underline{ao}_{<t}a_t). \tag{4.3}$$

The posterior probabilities $P(\theta|\underline{ao}_{<t}a_t)$ are calculated using Bayes' rule:

$$P(\theta|\underline{ao}_{<t}a_t) = \frac{P(o_{<t}|\theta, a_{\leq t})P(\theta)}{\sum_{\theta'} P(o_{<t}|\theta', a_{\leq t})P(\theta')} = \frac{P(\theta)\prod_{\tau=1}^{t-1} P(o_\tau|\theta, \underline{ao}_{<\tau})}{\sum_{\theta'} P(\theta')\prod_{\tau=1}^{t-1} P(o_\tau|\theta', \underline{ao}_{<\tau})}.$$

Here, we see that the trailing $a_t$ does not affect the observations $o_{<t}$ and thus can be dropped, i.e.

$$P(\theta|\underline{ao}_{<t}a_t) = P(\theta|\underline{ao}_{<t}). \tag{4.4}$$

Consider now without loss of generality the probability $P(\theta|\underline{ao}_{\leq t})$. Applying Bayes' rule over the last observation $o_t$, one obtains

$$P(\theta|\underline{ao}_{\leq t}) = \frac{P(o_t|\theta, \underline{ao}_{<t})\, P(\theta|\underline{ao}_{<t})}{\sum_{\theta'} P(o_t|\theta', \underline{ao}_{<t})\, P(\theta'|\underline{ao}_{<t})}. \tag{4.5}$$

Collecting (4.3), (4.4) and (4.5) yields the result. ■

In (4.2), the posterior probabilities $P(\theta|\underline{ao}_{\leq t})$ play the role of adaptive mixture weights. Also note that the result shows that $P(\theta|\underline{ao}_{<t}a_t) = P(\theta|\underline{ao}_{<t})$, i.e. the trailing $a_t$ does not affect the posterior weights. The advantage of Proposition 1 is that of concisely highlighting the sequential nature of belief updates.

### 4.2.3 Induced Input Model

We have seen that the Bayesian input model allows deriving a predictor (i.e. the predictive distribution) over the input stream. This predictor can serve as the input model for an adaptive autonomous system.

**Definition 19 (Induced Input Model)** The input model **P induced** by a Bayesian input model $P$ is defined as

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) := \sum_\theta P(\theta|\underline{ao}_{<t})P(o_t|\theta, \underline{ao}_{<t}a_t) \quad \text{for all } \underline{ao}_{<t} \in \mathcal{Z}^\diamond.$$

□

The intuitive interpretation of this definition is as follows. The agent does not know the environment $\mathbf{Q}$. Probabilistically, this fact is made precise by assuming that $\mathbf{Q}$ is going to be drawn with probability $P(\theta)$ from a set $\mathcal{Q} := \{\mathbf{Q}_\theta\}_{\theta\in\Theta}$ of possible output models of environments *before* the interaction starts. Following a Bayesian approach, this is modeled by equating

$$P(o_t|\theta, \underline{ao}_{<t}a_t) = \mathbf{Q}_\theta(o_t|\underline{ao}_{<t}a_t).$$

Hence, the uncertainty over the environment $\mathbf{Q}_\theta$ is translated into the uncertainty over $\theta$. Accordingly, an alternative way of looking at this situation is to think about an environment that secretly chooses its "parameter" $\theta \in \Theta$ before the interaction starts. Because of this, the set $\Theta$ is known as the set of **unknown parameters**.

**Remark 11** The distinction between the Bayesian input model and the input model, and the use of different probability symbols $P$ and $\mathbf{P}$ respectively, is meaningful, but this will only become clear later in Chapter 8 when we will introduce causal interventions.□

**Remark 12** While this setup models agents that are uncertain about its environment, it assumes that agents are certain about their own policy. In other words, the choice of the policy (causally) precedes the interactions. As we will see later, this is an assumption that cannot be met in most situations. □

### 4.2.4 Convergence of Predictive Distribution

So far, the usage of a Bayesian input model to describe the uncertainty in an autonomous system has been justified from a purely axiomatic point of view. However, the predictive distribution has an astonishing statistical property. If the Bayesian model contains the true input model (i.e. the input model describing the environment) then the predictive distribution converges to it.

**Theorem 3 (Convergence of Predictive Distribution)** *Let* $\mathbf{P}$ *be an input model induced by a Bayesian model* $P$*. Let* $\theta_* \in \Theta$ *be the true parameter, i.e.* $P(o_t|\theta_*, \underline{ao}_{<t}a_t) = \mathbf{Q}(o_t|\underline{ao}_{<t}a_t)$*. Let* $S$ *and* $D$ *defined as*

$$S := \sum_{t=1}^{T} \sum_{o_{\leq t}} \mathbf{Q}(o_{<t}|a_{<t}) \Big( \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) - \mathbf{P}(o_t|\underline{ao}_{<t}a_t) \Big)^2$$

$$D := \sum_{o_{\leq T}} \mathbf{Q}(o_{\leq T}|a_{\leq T}) \ln \frac{\mathbf{Q}(o_{\leq T}|a_{\leq T})}{\mathbf{P}(o_{\leq T}|a_{\leq T})}$$

*be the cumulative sum of the mean-squared prediction error and the relative entropy respectively. Then, the following inequalities hold:*

$$0 \leq S \leq D \leq \ln \frac{1}{P(\theta_*)}.$$

□

Proof This proof is due to Hutter (2004a). We want to apply the *entropy inequality* (Hutter, 2004a, Section 3.9.2)

$$\sum_i (x_i - y_i)^2 \leq \sum_i x_i \ln \frac{x_i}{y_i}, \quad \text{for} \quad x_i \geq 0, y_i \geq 0, \quad \sum_i x_i = 1, \sum_i y_i = 1.$$

In the entropy inequality, identify $x_i \leftrightarrow \mathbf{Q}(o_t|\underline{ao}_{<t}a_t)$, $y_i \leftrightarrow \mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ and carry out the sums over $o_t$,

$$\sum_{o_t} (\mathbf{Q}(o_t|\underline{ao}_{<t}a_t) - \mathbf{P}(o_t|\underline{ao}_{<t}a_t))^2 \leq \sum_{o_t} \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) \ln \frac{\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)}{\mathbf{P}(o_t|\underline{ao}_{<t}a_t)}.$$

We multiply the inequality by $\mathbf{Q}(o_{<t}|a_{<t})$ and take sums over the $o_{<t}$ and over the time steps $t = 1, \ldots, T$. We further apply the chain rule $\mathbf{Q}(o_{<t}|a_{<t})\mathbf{Q}(o_t|\underline{ao}_{<t}a_t) = \mathbf{Q}(o_{\leq t}|a_{\leq t})$ on the r.h.s. and obtain

$$\sum_{t=1}^{T}\sum_{o_{\leq t}}\mathbf{Q}(o_{<t}|a_{<t})\Big(\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)-\mathbf{P}(o_t|\underline{ao}_{<t}a_t)\Big)^2 \leq \sum_{t=1}^{T}\sum_{o_{\leq t}}\mathbf{Q}(o_{\leq t}|a_{\leq t})\ln\frac{\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)}{\mathbf{P}(o_t|\underline{ao}_{<t}a_t)}.$$

The r.h.s. can be further developed,

$$\begin{aligned}
&\overset{(a)}{=}\sum_{t=1}^{T}\sum_{o_{\leq t}}\mathbf{Q}(o_{\leq t}|a_{\leq t})\ln\frac{\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)}{\mathbf{P}(o_t|\underline{ao}_{<t}a_t)} \overset{(b)}{=}\sum_{o_{\leq T}}\mathbf{Q}(o_{\leq T}|a_{\leq T})\ln\prod_{t=1}^{T}\frac{\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)}{\mathbf{P}(o_t|\underline{ao}_{<t}a_t)}\\
&\overset{(c)}{=}\sum_{o_{\leq T}}\mathbf{Q}(o_{\leq T}|a_{\leq T})\ln\frac{\mathbf{Q}(o_{\leq T}|a_{\leq T})}{\mathbf{P}(o_{\leq T}|a_{\leq T})} \overset{(d)}{=}\sum_{o_{\leq T}}\mathbf{Q}(o_{\leq T}|a_{\leq T})\ln\frac{\mathbf{Q}(o_{\leq T}|a_{\leq T})}{\sum_\theta P(\theta)P(o_{\leq T}|\theta,a_{\leq T})}\\
&\overset{(e)}{\leq}\sum_{o_{\leq T}}\mathbf{Q}(o_{\leq T}|a_{\leq T})\ln\frac{\mathbf{Q}(o_{\leq T}|a_{\leq T})}{P(\theta_*)P(o_{\leq T}|\theta_*,a_{\leq T})} \overset{(f)}{=}\ln\frac{1}{P(\theta_*)}.
\end{aligned}$$

In (a), we changed $\sum_{o_{\leq t}}\mathbf{Q}(o_{\leq t}|a_{\leq t})$ to $\sum_{o_{\leq T}}\mathbf{Q}(o_{\leq T}|a_{\leq T})$. This can be done because the argument in the logarithm does not depend on $o_{t+1:T}$. (b) follows from moving the $\sum_t$ inside of the logarithm and thus converting it into a product $\prod_t$. (c) follows from applying the chain rule. (d) is obtained by applying (4.1), and (e) follows from dropping all the summands in the denominator but the one corresponding to the true parameter $\theta_*$. Equality (f) follows from $P(o_{\leq t}|\theta_*, a_{\leq t}) = \mathbf{Q}(o_{\leq t}|a_{\leq t})$ and from then simplifying the expression. Note that $S$, $D$ (as defined in the theorem statement) and $\ln\frac{1}{P(\theta_*)}$ are positive quantities. ■

The theorem says essentially two things. First, the cumulative sum of prediction errors (in the mean-square sense) $S$ is bounded by the (total) relative entropy of the predictive distribution from the environment. Absolute deviations like those in $S$ are intuitively easier to grasp (because of their geometrical interpretation) than relative deviations like those in $D$. Second, the relative entropy is bounded from above by a constant that is independent of $T$. The key property leading to this conclusion is that the mixture distribution **dominates** any of its hypotheses, i.e.

$$\mathbf{P}(o_{\leq T}|a_{\leq T}) = P(o_{\leq T}|a_{\leq T}) \geq \alpha\, P(o_{\leq T}|\theta, a_{\leq T}), \qquad \alpha > 0. \tag{4.6}$$

(In particular, in the Bayesian input model $\alpha = P(\theta)$.) The dominance factor $\alpha$ can be arbitrarily small as long as it stays constant. Hence, the upper bound $-\ln\alpha$ is valid even in the limit $T \to \infty$. Since $S$ contains an infinite sum of positive terms, one concludes:

**Corollary 1** *Under the same conditions as in Theorem 3, one has*

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) \xrightarrow{t\to\infty} \mathbf{Q}(o_t|\underline{ao}_{<t}a_t)$$

*with* $\mathbf{G}$*-probability one.* □

PROOF Let $z_1(\omega), z_2(\omega), \ldots$ be a sequence of real-valued random variables drawn according to a probability measure $\mu$. $z_t$ is said to converge for $t \to \infty$ in the mean sum to a real value $z_*$ iff

$$\sum_{t=1}^{\infty} \mathbf{E}[(z_t - z_*)^2] < \infty.$$

It is a well-known result that convergence in the mean sum implies convergence with probability one, i.e.

$$\mu\{\omega : z_t(\omega) \to z_*\} = 1.$$

From Theorem 3 we know that

$$\sum_{t=1}^{T} \sum_{o_{\leq t}} \mathbf{Q}(o_{<t}|a_{<t}) \Big( \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) - \mathbf{P}(o_t|\underline{ao}_{<t}a_t) \Big)^2 \leq -\ln P(\theta_*) < \infty.$$

In particular, it can be verified that this bound holds for the case $P(\theta_*) = 1$. Extending the sum over $o_{\leq t}$ to $o_{\leq T}$ and then averaging over the $a_{\leq T}$ preserves the inequality. This yields

$$\sum_{t=1}^{T} \sum_{\underline{ao}_{\leq T}} \mathbf{G}(\underline{ao}_{\leq T}) \Big( \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) - \mathbf{P}(o_t|\underline{ao}_{<t}a_t) \Big)^2 < \infty,$$

where $\mathbf{G}$ is the generative probability measure

$$\mathbf{G}(\underline{ao}_{\leq T}) = \prod_{t=1}^{T} \mathbf{P}(a_t|\underline{ao}_{<t}) \mathbf{Q}(o_t|\underline{ao}_{<t}a_t).$$

Replacing $\mathbf{G} \leftrightarrow \mu$, $(\mathbf{P}(o_t|\underline{ao}_{<t}a_t) - \mathbf{Q}(o_t|\underline{ao}_{<t}a_t)) \leftrightarrow z_t$ and $0 \leftrightarrow z_*$ and applying "in the mean sum $\Rightarrow$ with probability one" we obtain the result. ■

Hence, in the limit $T \to \infty$, a random sequence $a_1 o_1 a_2 o_2 \ldots$ drawn from the generative measure $\mathbf{G}$ will have the property that the predictive distribution $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ converges to the environment's output stream $\mathbf{Q}(o_t|\underline{ao}_{<t}a_t)$.

To illustrate the convergence result, consider the coin-toss prediction example from the beginning of this chapter. Assume a Bayesian model with two possible parameter settings, namely $\Theta := \{\frac{1}{4}, \frac{3}{4}\}$. Hence,

$$P(o_t = \mathrm{H}|\theta, o_{<t}) = P(o_t = \mathrm{H}|\theta) = \begin{cases} \frac{1}{4} & \text{if } \theta = \frac{1}{4}, \\ \frac{3}{4} & \text{if } \theta = \frac{3}{4}. \end{cases}$$

The two hypotheses are shown in Figure 4.7a and b. Next, assume a uniform prior distribution over the parameters, i.e.

$$P(\theta) = \frac{1}{2}, \qquad \theta = \frac{1}{4}, \frac{3}{4}.$$

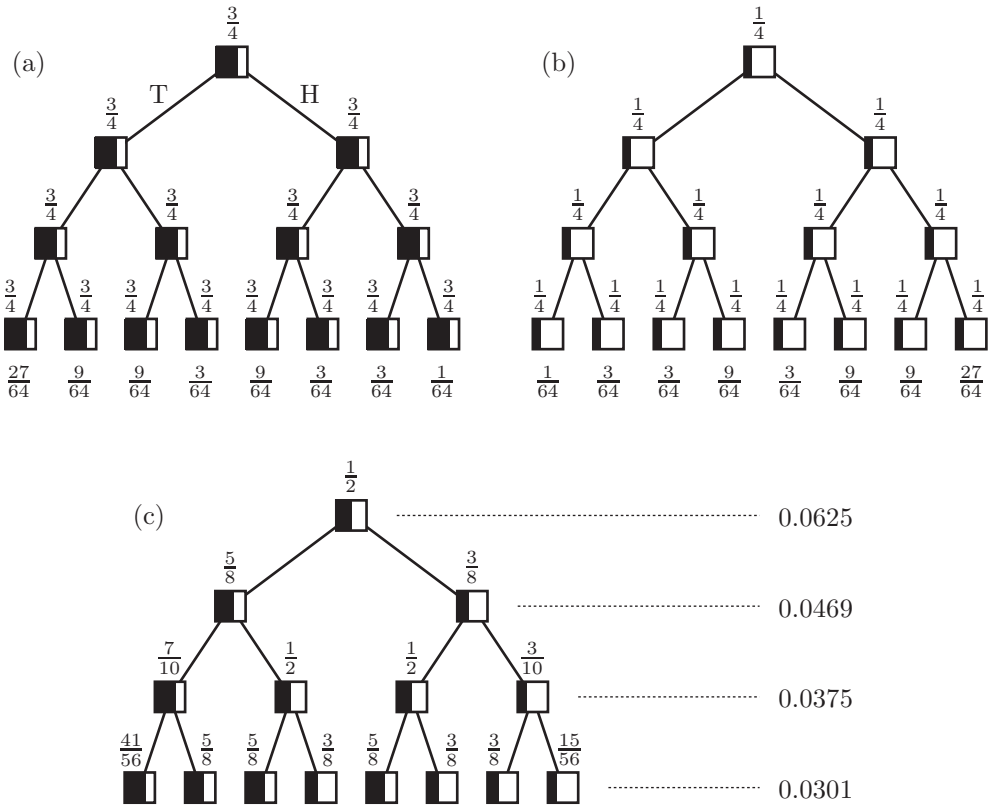



**Figure 4.7:** Convergence of Predictive Distribution. Panels (a) and (b) illustrate the hypotheses $\theta = \frac{1}{4}$ and $\theta = \frac{3}{4}$ respectively. The boxes represent the probabilities $P(o_t = \mathrm{T}|\theta, o_{<t})$ (black) and $P(o_t = \mathrm{H}|\theta, o_{<t})$ (white). The resulting probabilities of drawing a particular realization are indicated at the leaves. Panel (c) illustrates the predictive distribution $\mathbf{P}(o_t|o_{<t})$ resulting from combining the two hypotheses with prior probabilities $P(\theta = \frac{1}{4}) = P(\theta = \frac{3}{4}) = \frac{1}{2}$. Assume any of the two hypotheses is true. The numbers written at each level correspond to the mean-squared prediction error at time $t$, i.e. $\mathbf{E}[s_t]$, where $s_t := (\mathbf{Q}(o_t|o_{<t}) - \mathbf{P}(o_t|o_{<t}))^2$. (The $s_t$ are the same for both hypotheses because of the symmetry.) Note how this average monotonically decreases over time.

The posterior probability $P(\theta = \frac{1}{4}|o_{\leq t})$ is calculated using Bayes' rule:

$$P(\theta = \tfrac{1}{4}|o_{\leq t}) = \frac{P(o_{\leq t}|\theta = \frac{1}{4})P(\theta)}{\sum_{\theta'} P(o_{\leq t}|\theta')P(\theta')} = \frac{P(\theta)\prod_{\tau=1}^{t} P(o_\tau|\theta = \frac{1}{4}, o_{<\tau})}{\sum_{\theta'} P(\theta')\prod_{\tau=1}^{t} P(o_\tau|\theta', o_{<\tau})}$$
$$= \frac{(\frac{1}{2})\,(\frac{3}{4})^h\,(\frac{1}{4})^{t-h}}{(\frac{1}{2})\,(\frac{3}{4})^h\,(\frac{1}{4})^{t-h} + (\frac{1}{2})\,(\frac{3}{4})^{t-h}\,(\frac{1}{4})^h} = \frac{1}{1+3^{t-2h}},$$

where $h$ is the number of times a head has been observed in the first $t$ coin tosses. The posterior probability $P(\theta = \frac{3}{4}|o_{\leq t})$ is obtained following a similar calculation. This yields a posterior distribution given by

$$P(\theta = \tfrac{1}{4}|o_{\leq t}) = \frac{1}{1+3^{t-2h}} \qquad \text{and} \quad P(\theta = \tfrac{3}{4}|o_{\leq t}) = \frac{1}{1+3^{2h-t}},$$

The induced input model is the predictive distribution given by

$$\mathbf{P}(o_{t+1} = \mathrm{H}|o_{\leq t}) = \sum_{\theta} P(\theta|o_{\leq t})P(o_{t+1} = \mathrm{H}|\theta, o_{\leq t})$$
$$= \frac{1}{4}\cdot\frac{1}{1+3^{t-2h}} + \frac{3}{4}\cdot\frac{1}{1+3^{2h-t}}.$$

This predictor is shown in Figure 4.7c. In the figure, it is seen how the predictive distribution approaches the hypotheses that seems more plausible given the experience: paths from the root that have more heads reach a leaf with bias $\approx \frac{3}{4}$, and paths that have more tails reach a leaf with bias $\approx \frac{1}{4}$.

This is a simple predictor with only two hypotheses. If one increases the number of possible bias settings to the unit interval $\Theta := [0, 1]$ and places a uniform prior *density* $p(\theta) = 1$ over $\Theta$, then the resulting predictive distribution turns out to be the famous rule of succession

$$\mathbf{P}(o_{t+1} = \mathrm{H}|o_{\leq t}) = \frac{h+1}{t+2}$$

presented at the beginning of this chapter. Accordingly, this predictor converges to *any bias* $\theta_* \in \Theta$. Thus, some simple predictors implement infinite mixtures of predictors!

**Remark 13** Theorem 3 has additional implications. For instance, it provides a "convergence rate" in the sense that it bounds the expected amount of times the mean-squared error exceeds a given tolerance. □

**Remark 14** The convergence of the predictive distribution holds *for any policy*, since the actions $a_{\leq t}$ in Theorem 3 are arbitrary conditionals. □

**Remark 15** Note that there are no restrictions on the statistical properties of the hypotheses. For instance, the input models $P(o_t|\theta, \underline{ao}_{<t}a_t)$ are neither required to be i.i.d., stationary nor Markov. The convergence of the predictive distribution holds even with complex time-series. □

**Remark 16** The convergence of the posterior distribution to the true parameter $\theta_*$, i.e. $P(\theta|\underline{ao}_{<t}) \to \delta^{\theta}_{\theta_*}$ as $t \to \infty$, *does not hold in general*. This is simply because having two distinct parameters $\theta \neq \theta'$ does not imply that $P(\underline{ao}_{<t}|\theta) \neq P(\underline{ao}_{<t}|\theta')$, i.e. it does not imply that their likelihood functions are different under all realizations. For such a convergence result to hold, one has to make additional assumptions (e.g. ergodicity, i.i.d., etc.) □

**Remark 17** The key property for finding an upper bound in Theorem 3, namely *dominance* (Equation 4.6), is a very powerful property that holds even in a number of other situations where $\theta_* \notin \Theta$. □

### 4.2.5 Bayes Optimal Agents

The main question we have addressed in this section is: If the environment is unknown, how do we construct an optimal agent? The Bayesian model introduced above allows specifying a class of possible environments that the agent is uncertain about. This leads to the specification of prior probabilities $P(\theta)$ and conditional probabilities $P(o_t|\theta, \underline{ao}_{<t}a_t)$. As explained previously, defining these quantities completely determines the agent's input model $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$:

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) = \sum_{\theta} P(o_t|\theta, \underline{ao}_{<t}a_t)P(\theta|\underline{ao}_{<t}a_t).$$

*The important observation is that one can interpret this distribution as if it represented a "real" environment* $\tilde{\mathbf{Q}}$ *with the property that*

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) = \tilde{\mathbf{Q}}(o_t|\underline{ao}_{<t}a_t)$$

*for all* $\underline{ao}_{\leq t} \in \mathcal{Z}^{\diamond}$. This view allows the construction of a rational agent following the maximum SEU principle, e.g. by solving the Bellman optimality equations from Section 3.2.3. That is, given the input model $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ and a utility function $\mathbf{U}$ over $\mathcal{Z}^T$, choose a policy $\mathbf{P}(a_t|\underline{ao}_{<t})$ as

$$\mathbf{P}(a_t|\underline{ao}_{<t}) := \delta^{a_t^*}_{a_t}, \quad \text{where} \quad a_t^* := \arg\max_{a_t} \sum_{o_t} \mathbf{P}(o_t|\underline{ao}_{<t}a_t)F(\underline{ao}_{\leq t}),$$

where $F$ is the future utility function from Definition 12. The resulting policy is known as the **Bayes optimal policy**, and it constitutes the solution to the **adaptive optimal control problem**.

Let us recapitulate what we have achieved. The previous equation is an important result in control theory and artificial intelligence. It explains how to design adaptive autonomous systems: adaptive cannons, maze solvers and robots for Martian exploration—at least, in SEU theory!

## 4.3 Historical Remarks & References

As in the case of decision theory, probability theory has also had a long and controversial history, and in fact their development has always been closely connected. Currently, there are several schools of thought who defend different interpretations of probability. Roughly, they can be classified into *physical* and *evidential*[5].

The physical interpretation of probability sees probabilities as an intrinsic property of nature. These include the frequentist accounts (Venn, 1866; von Mises, 1919; Fisher, 1970; Neyman, 1950) and propensity accounts (Popper, 1934). In the frequentist interpretation, a probability corresponds to the relative frequency of the occurrence of an event in a random experiment that is repeated over time under similar conditions. In the propensity interpretation, a probability is a physical tendency (or chance) of an event to happen (even in a single trial).

The evidential interpretation sees probabilities as measures of degrees of belief in inference. These include the logical and epistemic accounts (Jeffreys, 1939; Cox, 1961; Jaynes and Bretthorst, 2003), personalistic (or "gambling") accounts (Ramsey, 1926; De Finetti, 1937; Savage, 1954), although the evidential interpretation can be traced back as far as to the 18th century (Bayes, 1763; Laplace, 1774). In the logical and epistemic accounts probabilities are extensions of plain truth values to partial truth values. In the personalistic interpretation, probabilities correspond to assumptions about the world that are necessary to justify decision making.

While the different schools of thought disagree on the interpretation of probabilities, they do agree on the axioms that govern them. These have been laid down in modern form by Kolmogorov (1933) following the success of measure theory (Borel, 1898; Lebesgue, 1904; Fréchet, 1915). The exposition presented in this chapter, which is an original contribution of this thesis, combines the basics of Kolmogorov (1933) and Jaynes and Bretthorst (2003).

Bayesian methods in control are as old as control theory itself. For instance, consider the work of Bellman (1957) and Kalman (1960). Currently, Bayesian methods are being applied extensively in diverse areas of control and reinforcement learning. Recent expositions on Bayesian methods in sequential decision making are for example Jordan, Ghahramani, and Saul (1997), Duff (2002), Hutter (2004a) and Legg (2008).

[5]They are sometimes also known as *objectivist* and *subjectivist* respectively in the literature.

# Chapter 5

# Problems of Classical Agency

So far, we have seen how to construct an optimal adaptive agent that is universal with respect to a given class of possible environments. This class can be very rich, containing a vast number of environments. For example, one could have a Bayesian input model spanning the set of *all computable environments*. Such a choice of the class of environments would lead to an agent that is able to learn any computable sequence: it will make predictions about the weather, play chess, solve mazes, drive cars, discover physical theories, solve IQ tests and predict future stock prices[1]. *While such an agent is possible in principle (within SEU theory), it is a fact that state-of-the-art implementations of adaptive agents do not even match fly intelligence.* There are many *technical* reasons for this lack of success, but there are also *theoretical* reasons that are deeply rooted in SEU theory. This is the subject of the second part of this thesis.

## 5.1 Computational Cost and Precedence of Policy Choice

According to the maximum SEU principle, the designer has a utility function $\mathbf{U}$ over $\mathcal{Z}^T$ and a predictor $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$, possibly derived from a Bayesian input model. The maximum SEU principle stipulates the choice of a policy $\mathbf{P}(a_t|\underline{ao}_{<t})$ maximizing the subjective expected utility

$$\sum_{\underline{ao}_{\leq T}} \mathbf{P}(\underline{ao}_{\leq T})\mathbf{U}(\underline{ao}_{\leq T}).$$

*For the maximum SEU principle to make sense, the choice of the optimal policy has to be made before the interaction starts, i.e. the policy has to be fully precomputed.*

What is not apparent from this simple formula, however, is that finding the optimal policy is very difficult safe for simple toy problems. Essentially, the number of candidate policies is too large, even if we consider only deterministic policies.

Recall from Section 3.2.1 that choosing a deterministic policy $\mathbf{P}(a_t|\underline{ao}_{<t})$ amounts to choosing a behavioral function $\pi$ from the set $\Pi$. This choice specifies the actions to

[1] These are all activities that are at least approximable with a computational method.

be chosen in each decision node. In the first level, there is only one decision node. In the second level, there are $|\mathcal{A}| \cdot |\mathcal{O}| = |\mathcal{Z}|$ decision nodes. Similarly, in level $t$, there are

$$|\mathcal{Z}|^{t-1}$$

decision nodes. Because of the chronological property of policies, the causal dependencies amongst the decision nodes are such that the decisions in the lower levels are independent from the decisions further up the decision tree—which is exactly the property that is exploited in order to formulate the Bellman optimality equations. Hence, the optimal policy is found by choosing the optimal action of each decision node at level $T-1$, then at level $T-2$, and so forth. Thus, if solving one decision node is one elementary operation, then the total number of operations $C$ is given by

$$C := |\mathcal{Z}|^{T-1} + |\mathcal{Z}|^{T-2} + \cdots + |\mathcal{Z}| + 1 = \frac{|\mathcal{Z}|^T - 1}{|\mathcal{Z}| - 1}.$$

As a matter of illustration, consider the construction of a very simple cybernetic system having only two possible inputs and outputs, i.e. $|\mathcal{A}| = |\mathcal{O}| = 2$, running for one minute at one action per second, i.e. a horizon of $T = 60$ time steps. (This is way simpler than the structural design of an ant.) Applying the maximum SEU principle requires

$$C = \frac{4^{60} - 1}{4 - 1} \approx 4.43 \cdot 10^{35}$$

operations—which is clearly an unreasonably large number considering that there are only about $10^{50}$ atoms in the world[2]!

*Hence, any rigorous application of the maximum SEU principle to choose an optimal policy must rely on severe domain assumptions that reduce the effective cardinality of the policy space to a manageable size.*

## 5.2 Is Rationality a Useful Concept?

The maximum SEU principle has always been advocated as a normative principle. However, given that there are virtually no real-world application domains due to its dramatic computational demands, it not unreasonable to question the theoretical usefulness of SEU theory and the notion of rationality all together. One can tackle this question in at least three ways.

1. *Rationality as the "gold standard".* One could easily imagine that computing the optimal answer is so costly, that one would rather content oneself with a "suboptimal" (but strictly speaking, irrational) solution that incurs into less resource costs. In this view, the maximum SEU principle is taken as a "gold standard" that has to be approximated. This is the view shared by most of the engineering community.

[2] And if this does not convince yet, consider that today's fastest supercomputer carrying out $10^{16}$ operations per second would take $1.40 \cdot 10^{10}$ years to find the optimal policy (which, ironically, lasts only one minute), which is even more than the age of the universe, being "only" about $1.37 \cdot 10^{10}$ years!

2. *Rationality as an idealization.* Scientific models are usually idealized descriptions of phenomena happening in Nature, and some models are empirically more accurate than others. The model of rationality is not different in this sense: it is an idealization that has the advantage of being mathematically simple and elegant. While it is inaccurate, it captures important aspects of the decision making process. In this view, rationality turns out to be useful concept to study many situations arising in real life situations. This is the view defended by many economists.

3. *Searching for new foundations.* If the framework of rationality fails to capture some aspects of decision making that are considered important (like the resource costs of finding the solution), then one can search for new foundations to remedy the shortcomings of SEU theory. In this view, the aim is to find a theory to conceptualize behavior under limited resources. This is the approach pursued by the paradigm of *bounded rationality*, as well as the goal of the second part of this thesis.

## 5.3 Historical Remarks & References

The distinction between decision theory as a normative and as a descriptive theory has been pointed out even during its first mathematical formalizations. For instance, see Luce (1959) for a different set of axioms. The need for considering computational resources in decision making has been first pointed out by Simon (1955). A modern reference to the field of bounded rationality is **?**.

# Part II

# Resource-Bounded Agency

# Chapter 6

# Resources

In the previous chapter we have argued that one of the major problems of building autonomous agents according to the maximum SEU principle is the prohibitive computational complexity of the method. This computational cost arises because the method requires us to find the optimal policy within a set of candidate policies that is way too large before the agent has even interacted once with the environment. It is therefore of crucial importance to formally understand this computational cost in order to propose methods to deal with it.

There are several ways to study computational costs in learning and in control. One widely developed field is **computational learning theory**, which studies the computational complexity and feasibility[1] of learning algorithms (Angluin, 1992; Kearns and Vazirani, 1994; Bishop, 2006). Another approach is **bounded rationality**, an approach to decision making that complements rational decision making by taking into account the cognitive/computational limitations of decision makers (Simon, 1955; Rubinstein, 1988). In particular, in artificial intelligence, bounded rationality is viewed as the ability to reason about the resource-costs of reasoning, i.e. **meta-reasoning** (Russell and Wefald, 1991). In this view, a resource-bounded agent must decide what to reason about, when, and for how long. This leads to agents that trade off computational expenses of reasoning against the expected utility of the outcome. However, this approach to bounded rationality is not devoid of criticism. One can easily envisage an infinite hierarchy of meta-problems of "reasoning-about-reasoning" with no principled way to stop the infinite regression (Parkes, 1997).

In this chapter, the two main questions that we want to be able to address with a formalization of resources are:

1. What are the resources spent by the agent during its execution, i.e. its interaction with the environment?

2. What are the resources spent by the designer for the construction of the agent?

As we will argue in this chapter, answering these two questions requires understanding the relationship between:

[1] A computation is considered feasible when it can be carried out in polynomial time.

- the intuitive concept of resources;
- (transformations of) probability distributions;
- computational resources such as computation time and computation space.

As of today, the formal relationship between probability distributions and computational resources is not understood, and in fact relates to many deep open questions in complexity theory. Fortunately, some progress into this direction can be made by restricting ourselves to understanding the relationship between resources and probability distributions by drawing ideas from thermodynamics and coding theory.

Consequently, the aim of this chapter is to introduce an information-theoretic (or more accurately, a thermodynamic) formalization of resources. While this formalization is admittedly non-standard from a control-theoretic point of view, it will prove to be fruitful to reinterpret the construction and the behavior of an autonomous system. In particular, it will allow to state a natural link between resources and probabilities, in a way such that behavior can be thought of as arising only from resource costs. Furthermore, we will sketch an intuitive but non-rigorous connection to computational complexity.

## 6.1 Preliminaries in Information Theory

As has been anticipated in the introduction, our primary focus will be to establish an information-theoretic formalization of resources. To understand this connection, it is necessary to introduce some basic concepts of information theory, the mathematical theory concerned with the quantification of information (Shannon, 1948; MacKay, 2003).

### 6.1.1 The Communication Problem

The basic problem of information theory is that of communicating a message or a choice from a sender to a receiver via a communication **channel**, i.e. any device that translates an input symbol into an output symbol, possibly with some noise. This setup covers a wide range of scenarios, such as a modem transmitting content to another modem via an optic fiber; a parent cell passing genetic information to its offspring cells via its chromosomes; a hard disk storing data, i.e. delivering the data to itself but at a later point in time; and so forth. The fundamental result of information theory is that all the properties of the channel can essentially be reduced to its **capacity**, that is, the limit to the amount of (noiseless) information that it can transmit each time it is used.

In a communication problem, a sender communicates a choice to a receiver over a communication channel (see Figure 6.1). This is formalized as follows. Let $\mathcal{U}$ and $\mathcal{X}$ be two finite sets, where $\mathcal{U}$ is the set of possible choices, and where $\mathcal{X}$ is the alphabet used by the channel. A channel is a device that receives an input symbol $x \in \mathcal{X}$ and produces an output symbol $y \in \mathcal{X}$ following a conditional distribution $P(y|x)$. Since $\mathcal{X}$ is typically assumed to be much smaller than $\mathcal{U}$, a choice $u \in \mathcal{U}$ is communicated to the

receiver by encoding it as a string $x_{\leq n} \in \mathcal{X}$ and then transmitted symbol by symbol[2]. These symbols are recovered as a (possibly corrupted) string $y_{\leq n} \in \mathcal{X}^*$ which is then decoded as the object $v \in \mathcal{U}$.


**Figure 6.1:** The communication problem consists in communicating a (noisy) choice of a target object $v \in \mathcal{V}$ via the selection of a source object $u \in \mathcal{U}$ using a communication channel. This is done by encoding $u$ as a string $x_{\leq n} \in \mathcal{X}^*$ that is then transmitted over the channel. The transmitted string is recovered as a string $y_{\leq n} \in \mathcal{X}^*$ that is then decoded as the target object $v$. The blocks highlighted in bold correspond to the processes that have to be designed by the engineer.

Most of the time, we will assume a binary alphabet for the channel, that is $\mathcal{X} = \{0, 1\}$, because the choice of an alphabet changes the length of the codes by at most a constant factor that depends only on the size of the alphabet. We can thus think of the communication problem as the design of binary codes that efficiently transmit a choice using a given binary communication channel. Due to the central role that codes play in information theory, the next subsection investigates their basic properties.

### 6.1.2 Codes

A code is a scheme to represent objects as binary strings. A badly chosen code can lead to inefficient or ambiguous representations. An inefficient code is a code that does not minimize the length of its codewords (measured in bits), and an ambiguous code is a code where the original object cannot be recovered uniquely anymore. In this section we investigate the property that a code has to possess in order to be both efficient and unambiguous. The exposition in this subsection follows closely the one presented in Grünwald (2007). We first require a definition of suitable binary codes.

**Definition 20 (Prefix Free)** Given a finite alphabet $\mathcal{X}$, a subset $\mathcal{P} \subset \mathcal{X}^*$ is called **prefix free** iff there are no distinct members $u, v \in \mathcal{P}$ such that $u$ is a prefix of $v$. □

For instance, for the binary alphabet $\{0, 1\}$, the sets

$$\mathcal{P}_1 = \{001, 01, 101, 11\}, \qquad \mathcal{P}_2 = \{00, 01, 10, 11\} \qquad \text{and} \qquad \mathcal{P}_3 = \{0, 10, 110, 111\}$$

are prefix free, but

$$\mathcal{P}_4 = \{01, 011, 1\}$$

[2]Recall that $\mathcal{X}^*$ is the set of all finite strings over the alphabet $\mathcal{X}$.

is not because 01 is a prefix of 011. In fact, $\mathcal{P}_4$ is even ambiguous: if we want to decode the string 011 we would not be able to tell whether it corresponds to the codeword 011 alone or to the concatenation of the codewords 01 and 1.

**Definition 21 (Prefix Code)** Given a finite set of objects $\mathcal{U}$, a **prefix code** for $\mathcal{U}$ is an injective function $\mathbf{c} : \mathcal{U} \to \mathcal{P}$, where $\mathcal{P} \subset \{0,1\}^*$ is a prefix free set of binary strings. For any object $u \in \mathcal{U}$, $\mathbf{c}(u)$ is called the **codeword** of $u$, and $\mathbf{l}(u)$ denotes the corresponding **codeword length**. □

Prefix codes are important because of two reasons. First, they can be uniquely decoded; and second, when a given codeword is scanned from left to right, the end of the codeword is detected instantaneously. Prefix codes can be represented with a binary tree, where codewords correspond to the paths starting from the root until a leave is reached (Figure 6.2).

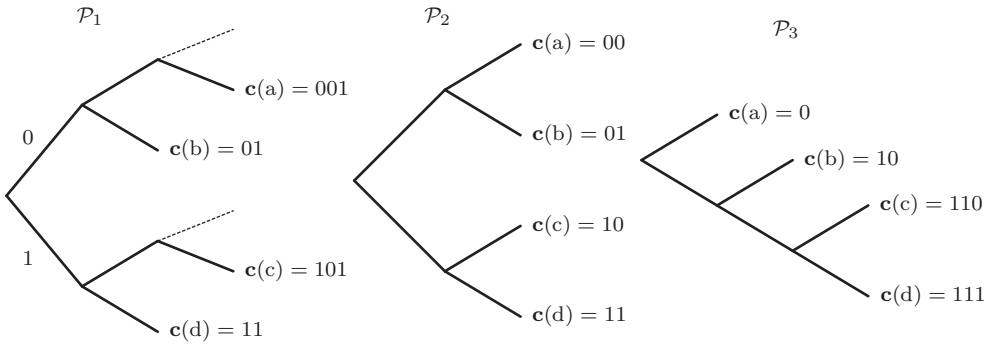



**Figure 6.2:** Three prefix codes $\mathcal{P}_1, \mathcal{P}_2$ and $\mathcal{P}_3$ over $\mathcal{U} = \{\text{a, b, c, d}\}$. Note that $\mathcal{P}_1$ could allocate codewords for at least two additional objects (namely, codewords with prefixes 000 or 100), while the latter ones cannot (i.e. they are complete). Furthermore, if one wants to shorten a codeword in a complete prefix code, then other codewords must necessarily grow.

From the examples in Figure 6.2, it is apparent that one cannot compress the codewords indefinitely. Codewords seem to allocate some "room", of which there is only a limited amount available. The shorter a codeword, the more "room" it takes away for allocating other codewords. For instance, compare the codes $\mathcal{P}_1$ and $\mathcal{P}_2$. In $\mathcal{P}_1$, the allocation of codewords is not optimal, in the sense that the codewords for 'a' and 'c' can still be shortened, as has been done in $\mathcal{P}_2$. Then, compare the codes $\mathcal{P}_2$ and $\mathcal{P}_3$. There, we decided to shorten the codeword for 'a' even more from 00 to 0. This however exceeds the limit, and hence in exchange we had to allocate longer codewords for the objects 'c' and 'd' to preserve the unique decodeability. This intuition is made precise by the following theorem.

**Theorem 4 (Kraft-McMillan Inequality)** *There is a prefix code over a finite set of objects $\mathcal{U}$ with codeword lengths $l_1, l_2, \ldots, l_{|\mathcal{U}|}$ iff they satisfy the* ***Kraft-McMillan***

***inequality***

$$\sum_{i=1}^{|\mathcal{U}|} 2^{-l_i} \leq 1.$$

*A prefix code whose codeword lengths satisfy this bound with equality is said to be* ***complete****.* □

PROOF ($\Rightarrow$) Interpret $0.u$ as the binary expansion of a real number in $[0, 1)$. For each $u \in \mathcal{U}$, let $\Gamma_u := [0.u, 0.u + 2^{-\mathbf{l}(u)})$, i.e. the right-open interval on the unit line starting at $0.u$ and ending at $0.u + 2^{-\mathbf{l}(u)}$. Observe that the measure of this interval is $2^{-\mathbf{l}(u)}$. Since the $\Gamma_u$ are disjoint and contained within $[0, 1)$, the total measure cannot exceed 1, thus proving that the inequality holds for prefix codes.

($\Leftarrow$) Assume w.l.g. that the $l_1, \ldots, l_{|\mathcal{U}|}$ are non-decreasing. Choose adjacent disjoint intervals $I_1, \ldots, I_{|\mathcal{U}|}$ of lengths $2^{-l_1}, \ldots, 2^{-l_{|\mathcal{U}|}}$ starting from the left end of the interval $[0, 1)$. The total measure of the union of the $I_i$ is $\leq 1$ by construction and the assumption about the $l_i$. Note that since the $l_i$ are non-decreasing, every interval $I_i$ ends up aligned with an interval $\Gamma_u$ for some binary string $u$. Take $u$ as the $i$-th codeword. ■

Thus, long codewords add little, while short codewords add much to the sum. Informally, one can say that long codewords are "cheap", while short codewords are "expensive". Accordingly, a prefix code that is incomplete can always be either compressed more or extended with extra codewords until it is complete.

But the Kraft-McMillan inequality allows us to say more than that. Consider a complete prefix code for $\mathcal{U}$. This code fulfills the equality:

$$\sum_{u \in \mathcal{U}} 2^{-\mathbf{l}(u)} = 1.$$

Now, *define*

$$\mathbf{Pr}(u) := 2^{-\mathbf{l}(u)} \qquad \text{for all } u \in \mathcal{U}.$$

We have obtained a probability distribution over $\mathcal{U}$! This is an important construction. Essentially, the Kraft-McMillan inequality allows establishing a one-to-one correspondence between codeword lengths and probabilities. This works even when we allow codeword lengths to take on any positive real values as long as they satisfy the Kraft-McMillan inequality. This enables us to talk about a codeword length as being a direct specification of a probability that is invariant to the particular choice of the encoding alphabet, and even to the actual codewords.

### 6.1.3 Information

We now return to the problem of communication, i.e. the problem of communicating a choice using a communication channel. In the previous subsection, we have seen that the Kraft-McMillan inequality characterizes the limits of compressibility for unambiguous

codes. We have also seen that the inequality allows us establishing a bijection between codeword lengths and probabilities given by

$$\mathbf{l}(u) = -\log \mathbf{Pr}(u), \tag{6.1}$$

where the quantity on the right-hand side is called the **information content** of $u$. The function $f(p) = -\log p$ is depicted in Figure 6.3a, where we use the convention $-\log 0 = \infty$. What is the operational meaning of this relation? Suppose that both the sender and the receiver agree on a code for $\mathcal{U}$, and that the communication channel is noiseless. When the sender chooses object $u \in \mathcal{U}$, he has to transmit $\mathbf{l}(u)$ bits through the channel. Assuming this choice is made following a probability distribution $\mathbf{Pr}$ over $\mathcal{U}$, the expected amount of transmitted bits is given by

$$\sum_u \mathbf{Pr}(u)\mathbf{l}(u).$$

If we want to make this communication as efficient as possible, then we have to choose a code that minimizes this expectation. One can show (MacKay, 2003, Section 5.3) that the optimal choice of the codeword lengths is precisely given by (6.1). Using (6.1), one sees that the minimum expected number of bits is given by

$$-\sum_u \mathbf{Pr}(u) \log \mathbf{Pr}(u)$$

which is called the **entropy** of the probability distribution $\mathbf{Pr}$. The function $f(p) = -p \log p$ is shown in Figure 6.3b, where we use the convention $0 \cdot \log 0 = 0$. The entropy can be shown to fulfill a series of properties that make it a suitable measure of the "uncertainty", "randomness" and the "average amount of information" contained in the choice of $u$ (Shannon, 1948; MacKay, 2003). However, if we choose a code with codeword lengths $\mathbf{l}'(u) \neq -\log \mathbf{Pr}(u)$, then the expected codeword length is the **cross-entropy**

$$\sum_u \mathbf{Pr}(u)\mathbf{l}'(u) = -\sum_u \mathbf{Pr}(u) \log \mathbf{Pr}'(u),$$

where $\mathbf{Pr}'(u) = 2^{-\mathbf{l}'(u)}$ are the probabilities "implicitly assumed" by the code. Obviously, the cross-entropy is lower bounded by the entropy,

$$-\sum_u \mathbf{Pr}(u) \log \mathbf{Pr}'(u) \geq -\sum_u \mathbf{Pr}(u) \log \mathbf{Pr}(u).$$

Notice how the efficiency of the communication does not only depend on the channel, but on its *usage* as well, i.e. the statistics of the information that is transmitted. Intuitively speaking, if we knew beforehand what choice was going to be made, then it would not be necessary to transmit any bit through the communication channel. But because we are *uncertain about the realization, we have to place bets*, bets which are implicitly captured by the code.

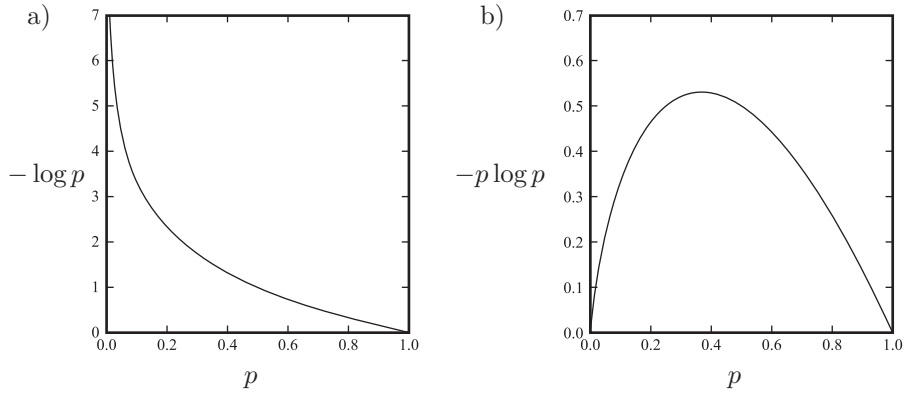



**Figure 6.3:** Information functions.

Thus, the information content corresponds to the amount of bits that have to be transmitted in order to communicate a choice. However, what happens *during* transmission, i.e. when not all the bits have been transmitted yet? The few bits that got sent did indeed communicate part of the choice. Communicating only part of the choice means that some of the possible choices were ruled out by the transmitted information. Consider the code $\mathcal{P}_3$ depicted in Figure 6.4a and assume that the first bit of the choice, say '1', is sent. This transmission changes the codeword function $\mathbf{c}$ to another codeword function $\mathbf{c}'$ depicted in Figure 6.4b, where the codewords starting with '1' are shortened by this bit and the codewords starting with '0' are discarded.

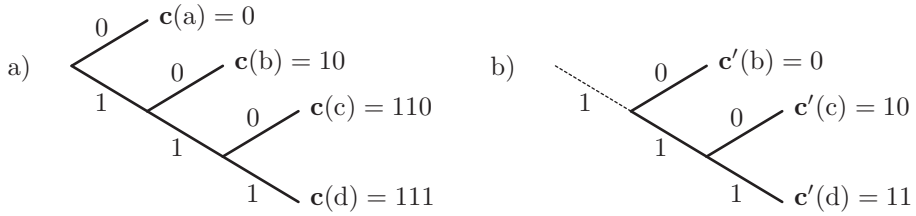



**Figure 6.4:** Probability versus Codeword Length (in bits).

The change in the expected codeword length can be calculated as

$$\left\{ \text{ final expected codeword length } \right\} - \left\{ \text{ initial expected codeword length } \right\}$$
$$= -\sum_u \mathbf{Pr}'(u) \log \mathbf{Pr}'(u) + \sum_u \mathbf{Pr}'(u) \log \mathbf{Pr}(u) = \sum_u \mathbf{Pr}'(u) \log \frac{\mathbf{Pr}(u)}{\mathbf{Pr}'(u)}.$$

In the previous calculation, the discarded codeword lengths are assumed to be equal to $\infty$ by convention, and the actual distribution is $\mathbf{Pr}'(u) = 2^{-\mathbf{l}'(u)}$, where $\mathbf{l}'$ is the

codeword length function associated to $\mathbf{c}'$. In the example of Figure 6.4, this change is

$$\sum_u \mathbf{Pr}'(u) \log \frac{\mathbf{Pr}(u)}{\mathbf{Pr}'(u)} = 0 \cdot (\infty - 0) + \frac{1}{2} \cdot (1 - 2) + \frac{1}{4} \cdot (1 - 3) + \frac{1}{4} \cdot (1 - 3) = -1,$$

as expected. Hence, the negative of the previous quantity, that is

$$\sum_u \mathbf{Pr}'(u) \log \frac{\mathbf{Pr}'(u)}{\mathbf{Pr}(u)},$$

called the **relative entropy** or **Kullback-Leibler divergence**, measures the amount of bits that have to be transmitted in order to change the knowledge about the choice from $\mathbf{Pr}(u)$ to $\mathbf{Pr}'(u)$. Note that the relative entropy is always positive and zero iff $\mathbf{Pr} = \mathbf{Pr}'$. *More generally, the relative entropy corresponds to the amount of bits that have to be paid in order to transform an initial probability distribution* $\mathbf{Pr}$ *into a final probability distribution* $\mathbf{Pr}'$. Because of this, the relative entropy will play a fundamental role in our formalization of resources.

## 6.2 Resources as Information

How does information theory help us to formalize resources in agents? In a communication problem, a natural way of measuring resources consists in counting the number of bits that are necessary to communicate a choice. The resource costs conceptualized in this way arise due to the uncertainty one has about the realization before it has happened. The setup of the communication problem is more general that it seems: given an appropriate interpretation, it can be related to thermodynamical and computational formalizations of resources.

### 6.2.1 Thermodynamical Interpretation

From a physical point of view, resources are typically expressed in terms of energy units (e.g. Joules or Calories). In particular, energy is defined via work, i.e. the amount of mechanical effort carried out on a system in order to change its physical state (Goldstein, 1980; Callen, 1985). For example, when an ideal gas is compressed with a piston under isothermal conditions from an initial volume $V$ to a final volume $V'$, then the work is calculated as

$$W = -\int_V^{V'} \frac{NRT}{V}\, dV = NRT \ln \frac{V}{V'}, \tag{6.2}$$

where $N \geq 0$ is the amount of substance, $R > 0$ is the gas constant, and $T \geq 0$ is the absolute temperature. The minus sign is just a convention to denote work done by the piston rather than by the gas. The interpretation of resources that we want to put forward here is analogous to the physical concept of work.

One can postulate a formal correspondence between one unit of information and one unit of work (Feynman, 2000). Consider representing one bit of information using

one of the following logical devices: a molecule that can be located either on the top or the bottom part of a box; a coin whose face-up side can be either head or tail; a door that can be either open or closed; a train that can be orientated facing either north or south; and so forth. Assume that all these devices are initialized in an undetermined logical state, where the first state has probability $p$ and the second probability $1 - p$. Now, imagine you want to set these devices to their first logical state. In the case of the molecule in a box, this means the following. Initially, the molecule is uniformly moving around within a space confined by two pistons as depicted in Figure 6.5a. Assuming that the initial volume is $V$, the molecule has to be pushed by the lower piston into the upper part of the box having volume $V' = pV$ (Figure 6.5b). From information theory, we know that the number of bits that we fix by this operation is given by

$$-\log p. \tag{6.3}$$

Using the formula in (6.2), one gets that the amount of work done by the piston is given by

$$W = RT \ln \frac{V}{V'} = RT \ln \frac{V}{pV} = -RT \ln p = -\frac{RT}{\log e} \log p = -\gamma_{\text{mol}} \log p,$$

where we have assumed $N = 1$ and where the constant $\gamma_{\text{mol}} := \frac{RT}{\log e} > 0$ can be interpreted as the conversion factor between one unit of information and one unit of work for the molecule-in-a-box device.

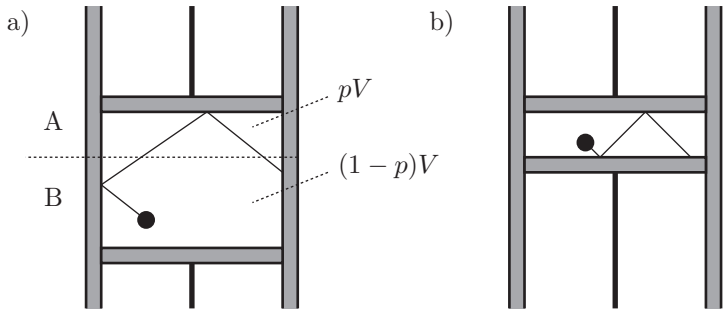



**Figure 6.5:** The Molecule-In-A-Box Device. (a) Initially, the molecule moves freely within a space of volume $V$ delimited by two pistons. The compartments A and B correspond to the two logical states of the device. (b) Then, the lower piston pushes the molecule into part A having volume $V' = pV$.

How do we compute the information and work for the case of the coin, door and train devices? The important observation is that we can model these cases as if they were like molecule-in-a-box devices, with the difference that their conversion factors between units of information and units of work are different. Hence, the number of bits fixed while these devices are set to the first state is given by

$$-\log p,$$

i.e. exactly as in the case of the molecule. However, the work is given by

$$-\gamma_{\text{coin}} \log p, \quad -\gamma_{\text{door}} \log p, \quad \text{and} \quad -\gamma_{\text{train}} \log p$$

respectively, where $\gamma_{\text{coin}}$, $\gamma_{\text{door}}$ and $\gamma_{\text{train}}$ are the associated conversion factors between units of information. Obviously,

$$\gamma_{\text{mol}} \leq \gamma_{\text{coin}} \leq \gamma_{\text{door}} \leq \gamma_{\text{train}}.$$

The point is that information is proportional to work. In other words, the amount of bits required to communicate a choice is proportional to the amount of thermodynamical work that has to be carried out on the recording device of the receiver.

One can easily envisage devices that generalize this idea to multiple states. Consider for instance the device shown in Figure 6.6a corresponding to the codeword lengths shown in Figure 6.4a. Here, four pistons control the probability of the molecule being in the parts of the space labeled as {a, b, c, d}. A choice is ruled out by pushing its piston downwards until it hits the wall. To rule out option 'a' as we did in the example of Figure 6.4, we push the first piston to the end in order to obtain the configuration shown in Figure 6.6b, which corresponds to codeword lengths of Figure 6.4b. Since this change reduces the total volume by half, the thermodynamical work is proportional to 1 bit.

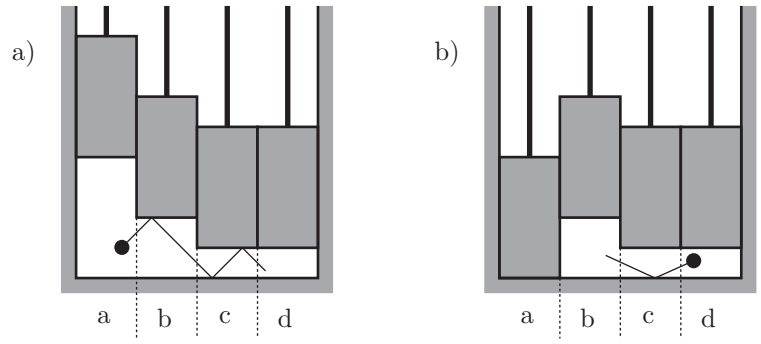



**Figure 6.6:** A generalized molecule-in-a-box device representing the partial choice of Figure 6.4. A molecule can move freely within a volume delimited by four pistons. The box has four sections labeled as 'a', 'b', 'c' and 'd', corresponding to four possible states of the device. Panel (a) and (b) show the initial and final configuration, corresponding to ruling out option 'a'.

In general, this calculation is done using the relative entropy. Let $v(u)$ and $v'(u)$ denote the volume of choice $u$ before and after the compression respectively, and let $\mathbf{Pr}(u)$ and $\mathbf{Pr}'(u)$ denote their associates probabilities. These probabilities are given by

$$\mathbf{Pr}(u) = \frac{v(u)}{V} \qquad \text{and} \qquad \mathbf{Pr}'(u) = \frac{v'(u)}{V'}.$$

Substituting the probabilities into the relative entropy, one obtains

$$\sum_u \mathbf{Pr}'(u) \log \frac{\mathbf{Pr}'(u)}{\mathbf{Pr}(u)} = \sum_u \frac{v'(u)}{V'} \log \frac{v(u)}{v'(u)} + \log \frac{V'}{V}.$$

Using the quantities of the example, one obtains

$$0 \cdot \log 0 + \frac{1}{8} \cdot \log 1 + \frac{1}{16} \cdot \log 1 + \frac{1}{16} \cdot \log 1 + \log 2 = 1 \text{ bit},$$

as expected. Note that this calculation only works if the volume stays equal or is compressed, but not expanded.

### 6.2.2 Computational Interpretation

**Remark 18** This subsection is of speculative nature. □

Usually, the efficiency of an algorithm is assessed in terms of a relevant computational resource. There are many possible computational resources, but the most important ones are the computation time and the computation space, i.e. the maximum number of time steps and the maximum number of cells that a Turing machine uses in order to compute the output of a function for any input. Typically, the goal of the designer is to construct an algorithm that minimizes either time or space.

However, these two goals seem to be somewhat incompatible. First, concentrating on a single resource does not represent all the issues involved in solving a problem. Also, there are numerous cases where reducing the computation time increases the computation space and viceversa. Indeed, the works of Borodin and Cook (1980), Beame (1989), Beame, Jayram, and Saks (2001) and Beame, Saks, Sun, and Vee (2003) derive tradeoff lower bounds for various problems such as sorting, finding unique elements in a set and solving randomized decision problems. Their findings show that *these lower bounds can be stated in terms of minimum time-space products*. For example, in Borodin and Cook (1980), an optimal $\Omega(n^2/\log n)$ lower bound is shown for the time-space product of sorting $n$ integers in the range of $[1, n^2]$. Although these findings are not conclusive, they seem to suggest that the time-space product might be a more general measure of computational resources, i.e. one that captures the intuitive notion of the *difficulty* of a computation (Figure 6.7). If one assumes that the time-space product is a suitable measure of computational resources, then one is led to ask about the meaning of this quantity.

Fortunately, one can give this product an information-theoretic meaning. We closely follow the presentation of Savage (1998). Assume we want to build a device that computes a function $f$ mapping $\mathcal{X}$ into $\mathcal{Y}$, where both $\mathcal{X}$ and $\mathcal{Y}$ are finite. To make our discussion concrete, suppose we want to implement this device using a **logic circuit**, i.e. a collection of interconnected logical gates computing elementary boolean functions. It is sufficient to restrict ourselves to binary AND, binary OR and unary NOT gates, since it can be shown that every boolean function can be implemented by them. An example circuit is shown in Figure 6.8. The complexity of a function is measured by

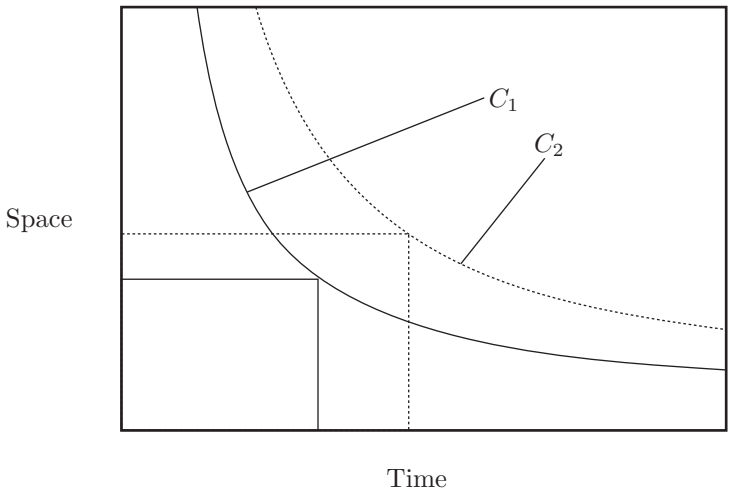



**Figure 6.7:** Time-Space Tradeoff. Studies suggest that there is a tradeoff between time and space that can be characterized in terms of a constant that acts as a lower bound on the time-space product of a problem. In the plot, the time-space curves of two problems $C_1$ and $C_2$ are shown. In this case, $C_2$ is more difficult than $C_1$ because the time-space product of $C_2$ is greater than the one of $C_1$, which can be seen by comparing their respective time-space rectangular areas.

counting the number of logic gates it has[3]. A function requiring more logic gates is considered to be more complex.

Many times we can also implement $f$ using a **sequential processing machine** which uses less logic gates. Such a machine is illustrated in Figure 6.9a. A sequential processing machine is capable of simulating computation models having a finite number of configurations, such as space-bounded **Turing machines** or space-bounded **random access machines** (Savage, 1998). As shown in the figure, a sequential processing machine is a logic circuit $\phi$ communicating with an external storage device (or "memory") $M$. The computation proceeds in $T$ discrete steps, where in step $t$, the logic circuit takes the binary input $x_t$ and the current state $q_t$ and generates the binary output $y_t$ and the next state $q_{t+1}$. This is done with the help of $M$, which stores the state $q_{t+1}$ generated at step $t$ until it is released in step $t+1$. Here, the input string $x_{1:T}$ encodes the input and the output string $y_{1:T}$ encodes the output.

The computation of a sequential processing machine can be rephrased as a communication problem. This is easily seen by "unwinding" the computation as shown in Figure 6.9, obtaining a concatenation of $T$ times the logic circuit $\phi$. This construction transforms the computation into a communication problem where the channel is given by the logic circuit $\phi$, that is, $\phi$ can be regarded as a noisy channel transforming $(x_t, q_t)$ into $(y_t, q_{t+1})$. If the state $q_t$ is represented by $S$ bits, then the total number of bits

[3]This measure is known as the circuit complexity of a function.

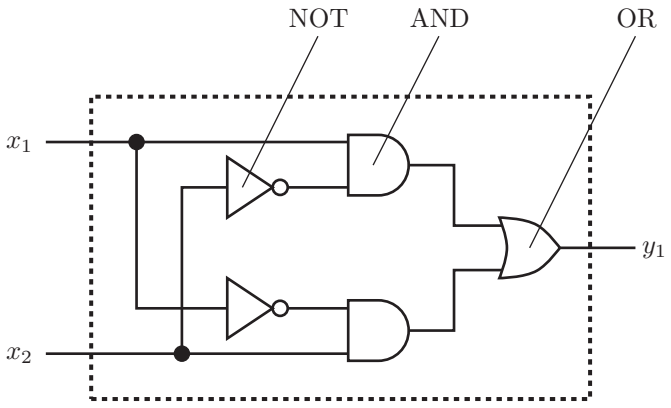



**Figure 6.8:** A logic circuit implementing the XOR function.

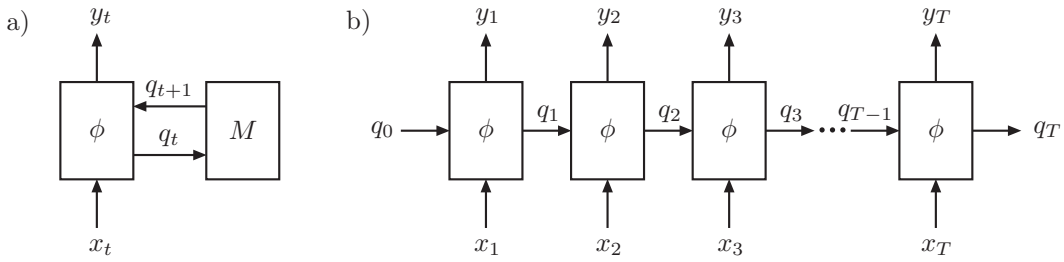



**Figure 6.9:** (a) A sequential processing machine consists of an logic circuit $\phi$ and an external storage device $M$. In each step $t$, $\phi$ takes the $t$-th input $x_t$ and state $q_t$ (stored in the external memory) and computes the $t$-th output $y_t$ and the $(t+1)$-th state $q_{t+1}$. (b) The computation of the sequential processing machine can be "unwinded", thereby constructing a large logic circuit consisting of $T$ concatenations of $\phi$ that computes the same function as before.

transmitted is given by the product

$$T \cdot (S + 1),$$

where it is seen that $(S + 1)$ is a measure of the maximum capacity that the circuit $\phi$, seen as a communication channel, can achieve. Note that $T$ and $S$ correspond to the time and space complexity of the function $f$. Let $C(f)$ and $C(\phi)$ denote the minimum number of logic gates needed to implement the functions $f$ and $\phi$ respectively. Then,

$$C(f) \leq T \cdot C(\phi) = \kappa \cdot T \cdot S,$$

because $C(\phi) = \kappa S$ for some positive $\kappa$ and because the implementation of $f$ as a sequential processing machine cannot use less logic gates than the direct implementation of $f$. Hence, the order of the time-space product is lower-bounded by the circuit complexity of $f$. In some sense, it seems that:

| Computer Science | | Information Theory |
|---|---|---|
| Compute $f(x)$ | $\Longleftrightarrow$ | Communicate $x$ and $f$ |

although the author is unaware of any proof of this claim.

The point is that, if we are willing to accept the time-space product as an appropriate measure of the computational complexity, then the computational resources can be related to information-theoretic resources—and as such, they are governed by information-theoretic principles. In other words, the computational resources correspond to the amount of bits that have to be transmitted to communicate a "computation". While this interpretation is intuitively appealing, there are still many open questions left. For instance, it is unclear what "computation" means in this sense, and it is also unclear how much computation is necessary in order to compute a given function (Sipser, 1996; Papadimitriou, 1993).

## 6.3 Resource Costs in Agents

In the previous section, we have seen that resources can be formalized in information-theoretic terms, and that this formalization can be given a thermodynamic and a computational interpretation. This can be summarized as follows. Given a set $\mathcal{U}$ of options, a receiver that does not know beforehand what the choice will be has to allocate resource costs (codeword lengths) that implicitly specify his beliefs $\mathbf{Pr}(u)$ about the realization of the choice $u \in \mathcal{U}$. Furthermore, when the receiver makes a (possibly stochastic and/or partial) choice represented by a probability distribution $\mathbf{Pr}'(u)$ over $\mathcal{U}$, then the amount of bits spent by the receiver in order to record this choice is given by the relative entropy

$$\sum_u \mathbf{Pr}'(u) \log \frac{\mathbf{Pr}'(u)}{\mathbf{Pr}(u)},$$

which is a generalization of the information content from deterministic choices to probabilistic choices. We have argued that this quantity is proportional to the amount of

thermodynamic work that the receiver has to carry out in order to record the choice. Furthermore, we have sketched a relation between the amount of bits and the required time-space product of the associated computation.

We argue that this way of thinking has many advantages: It greatly simplifies the analysis of resource costs, because we only have to deal with changes in probability distributions. This allows us abstracting away from the algorithmic details in order to reason about computational costs. The objective of this section is to explain how this formalization can be used to calculate the resource costs of running and constructing an agent.

### 6.3.1 Cost of Interaction

We have formalized autonomous systems as probability distributions over interaction sequences. According to the information-theoretic arguments presented in this chapter, this means that we are implicitly assigning resource costs for interactions.

This makes sense from an intuitive point of view. The implementation (or embodiment) of an agent facilitates some interactions while it hampers others. For instance, biologists can infer a great deal of the habits of a species by studying its anatomy. The rationale behind this is that animals manifest energy-efficient behavior more frequently than energy-intensive behavior. This kind of reasoning acquires an extreme form in paleontology, where behavior is mainly inferred from fossilized animals. Conversely, in engineering, systems are designed such that they minimize the resource costs of frequent or desirable operations and uses. This is visible in the designs of cars, aeroplanes, buildings, algorithms, advertising campaigns, etc.

From an information-theoretic point of view, an agent interacting with an environment is communicating an interaction sequence. Whenever the agent interacts with its environment (either producing an output or reading an input), its "physical state" or "internal configuration" changes as a necessary consequence of the interaction—simply because the two instants are distinguishable. In other words, if the instants before and after the interaction cannot be told apart, then we are forced to conclude that they are empirically the same. This change in "physical state" or "internal configuration" can take place in many possible ways: for instance, by a chemical reaction; by updating the internal memory; by consulting a random number generator; by moving to another location; or even by simply advancing in time (i.e. by changing the value of the time-coordinate). Hence, in this context, the semantics of "physical state" or "internal configuration" corresponds to an abstract information state: it is a description that exhaustively characterizes a situation of an agent. We call such a description a **state** of the agent.

To make this notion of states concrete, we introduce the following model. We assume that a change in state occurs whenever the agent either issues an action or reads an input. We start out from a blank binary tape that we will use to record the agent's experience. Then, we iteratively append a new binary string that encodes the new input or output symbol experienced by the agent (Figure 6.10). In this model, the appended binary strings are proxies for the changes in state that the agent experiences

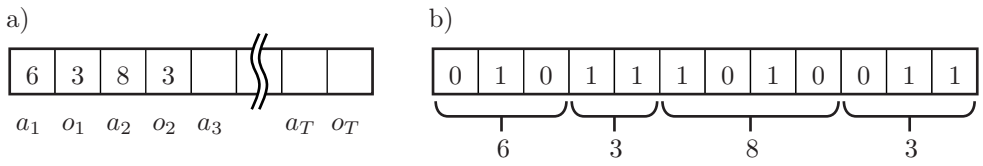



**Figure 6.10:** State Model. The agent has an I/O domain given by $\mathcal{A} := \mathcal{O} := \{1, 2, \ldots, 10\}$. So far, the agent has experienced four I/O symbols (Panel a). The state of the agent is constructed by iteratively encoding the four I/O symbols into binary strings (Panel b).

during its interactions with the environment. In this way, we abstract away from the inner workings of the agent by simply representing every change by a binary string. Note that this scheme does not allow an agent returning to a previous state, hence the agent cannot "jump back in time". Additionally, we want the content of the binary tape to be uniquely decodeable, such that we can recover the whole I/O history the agent has experienced so far at any given time by decoding the content of the tape.

This model highlights the correspondence between resources, codeword lengths and probabilities. Therefore, *the behavior of the agent can be thought of as a reflection of the underlying resource costs.*

### 6.3.2 Costs of Construction

**Remark 19** This subsection is of speculative nature. □

Designing and constructing an agent has a cost. Whether we are thinking to find a solution to optimality equations, running a search algorithm, or assembling mechatronic parts, we are always spending resources during the conception of an agent. These resource costs can be thought of as arising from the change of the distribution over the possible agents, where the cost of this change is given by the relative entropy.

Let $\Theta$ denote the index set parameterizing the possible agents. Furthermore, let $\mathbf{Pr}(\theta)$ denote the belief we have about $\theta$ being the optimal parameter. Consider the relative entropy

$$\rho = \sum_{\theta} \mathbf{Pr}'(\theta) \log \frac{\mathbf{Pr}'(\theta)}{\mathbf{Pr}(\theta)}, \tag{6.4}$$

where $\mathbf{Pr}'(\theta)$ is the (partial) choice made by the sender. This quantity correctly measures the number of bits we are going to receive over a noiseless channel from a sender that picks out the right $\theta$.

However, there is caveat: The previous calculation represents a situation where we passively receive the optimal answer, which is not possible unless the sender is an "oracle" who guesses the first $\rho$ bits of the optimal $\theta$. To correctly calculate the required number of bits in this communication problem, we have to account for the fact that *we do not know the sender*. Not knowing the sender has important implications, since this uncertainty might lead to significant resource costs that we are not accounting for.

Unfortunately, the author is not aware of any widely accepted way of dealing with this situation. However, one can *speculate* that the amount of information is

$$2^{\mathrm{O}(\rho)},$$

that is, exponential in the amount of information needed when the sender is known. Roughly speaking, the justification for this intuition is based on results from computational complexity, where a non-deterministic machine can be simulated by deterministic machine but incurring exponential cost (Sipser, 1996; Papadimitriou, 1993). In the next chapter, this claim is made precise for a special case.

## 6.4 Historical Remarks & References

The fundamental results of information theory were almost entirely developed in the paper by Shannon (1948). In particular, Shannon's paper derives a quantitative measure of information based on three desiderata. Surprisingly, the resulting formula for information turned out to have the same mathematical shape as the formula for thermodynamic entropy discovered empirically by Boltzmann (in a simpler form) and Gibbs. Because of this, much of information theory borrows mathematics from thermodynamics. Information theory has found a wide range of applications in communication, coding, compression, statistical learning, dynamical systems, and other areas (Khinchin, 1957; Ash, 1965; Kolmogorov, 1968; Gallager, 1968; Billingsley, 1978; Cover and Thomas, 1991; Li and Vitanyi, 2008; MacKay, 2003; Grünwald, 2007). The relation between information (more specifically, codeword lengths) and probability, follows roughly the argument presented in Grünwald (2007). The Kraft-McMillan inequality was developed in two steps by Kraft (1949) and then by McMillan (1956).

The relative entropy

$$\sum_u \mathbf{Pr}'(u) \log \frac{\mathbf{Pr}'(u)}{\mathbf{Pr}(u)}$$

has been introduced by Kullback and Leibler (1951). The standard interpretation is as follows. The relative entropy measures the expected number of extra bits required to code samples from $\mathbf{Pr}'$ when using a code based on $\mathbf{Pr}$, rather than using a code based on $\mathbf{Pr}'$. The idea of the relative entropy as a generalized formula for the information content is the author's (non-standard) interpretation. While mathematically equivalent, conceptually the author's interpretation seems to suggest a "temporal directionality", where $\mathbf{Pr}$ and $\mathbf{Pr}'$ represent the knowledge state of the receiver before and after receiving information.

The connections between information theory, thermodynamics and computational complexity presented in this chapter borrow ideas from various sources. The relation between units of energy and units of information, has been originally put forward in the context of Maxwell's demon by various authors (Maxwell, 1867; Szilard, 1929; Brillouin, 1951, 1956; Gabor, 1964). The first modern argument linking information theory with thermodynamics is due to Landauer (1961). Since then, the ideas of the *thermodynamics of computing* have found wide acceptance (Tribus and McIrvine, 1971; Bennett, 1973, 1982; Feynman, 2000; Li and Vitanyi, 2008). The devices shown in figures 6.5 and 6.6 are the author's contribution, and they differ from the devices in the literature in that they do not allow expanding the volume (i.e. erasing knowledge) but only compressing it (i.e. acquiring knowledge).

The relation between computational resources and information is an area under active research. Time-space tradeoffs have been investigated in Borodin and Cook (1980), Beame (1989),

Beame et al. (2001) and Beame et al. (2003) using a computational model called **branching programs**. The time-space tradeoff related to circuit complexity is presented in Savage (1998). The speculations relating the transmission of information to computational complexity are due to the author, although related ideas have been put forward by Bremermann (1965).

The arguments linking information-theoretic resources to the cost of interaction and construction are an original contribution of the author. The concept of an oracle, while non-standard in the context of information theory, is commonplace in computational complexity. An oracle is a hypothetical device which is capable of answering decision problems in a single operation (Sipser, 1996; Papadimitriou, 1993), and they are widely used in cryptography to make arguments about the security of cryptographic protocols involving hash functions.

# Chapter 7

# Boundedness

If a designer has unlimited computational resources, then he would pay the cost of calculating the optimal policy for an autonomous agent. However, in Chapter 5 we have argued that the rigorous application of the maximum SEU principle is computationally too expensive to serve as a practical design principle. Therefore, most implementations make severe domain restrictions and/or approximations in order to significantly reduce the computational expenses. In other words, designers content themselves with suboptimal solutions to the original problem.

The preference of the designer of choosing a suboptimal solution over the optimal one suggests that resources have an impact over the "perceived utility" of the solution. One can argue that this "suboptimal solution" becomes the "optimal solution" when resources are taken into account. Hence, intuitively there seems to be a common "currency" for resources and utilities. What is this currency? If there were one, then one could compute the optimal solution that trades off the benefits obtained from maximizing the expected subjective utility and the resource costs of the calculation.

## 7.1 An Example of Boundedness

We start our discussion with a concrete example. We are given an array of $N$ numbers $(v_1, v_2, \ldots, v_N)$, and our task is to pick the largest one. We solve this problem by checking the numbers one by one in any order, always comparing the current value against the largest seen so far. It is easy to see that this algorithm takes $\mathrm{O}(N)$ time and $\mathrm{O}(\log N)$ space if we assume that each comparison is done in a single computational step and that the indices are represented using $\lceil \log N \rceil$ bits. The time-space complexity of this algorithm is $\mathrm{O}(N \log N)$.

From an information-theoretic point of view, this algorithm transforms the knowledge state about the maximum. Figure 7.1 shows an example array having $N = 10$ elements with values in $\{1, 2, \ldots, 10\}$. Initially, the algorithm doesn't know the location of the maximum: If we assume that each index is encoded with $\log 10 \approx 3.3219$ bits, then the initial distribution is uniform. After running the algorithm, the location of the maximum is known, which is represented by a delta function concentrating its

probability mass on number 10. Notice that for discovering $\log N$ bits of information we are running an algorithm of time complexity $\mathrm{O}(N) = 2^{\mathrm{O}(\log N)}$, that is, exponential in the number of bits. Compare this to the arguments for the cost of construction presented in Section 6.3.2, page 68.

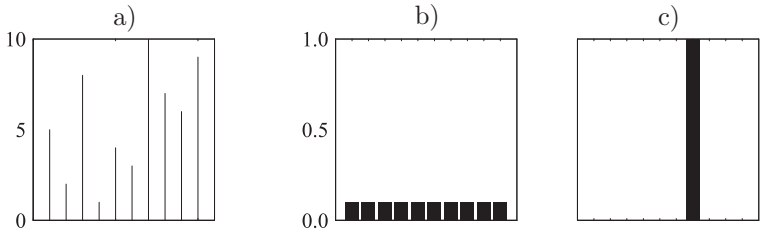



**Figure 7.1:** An Exhaustive Optimization. (a) A shuffled array with numbers from 1 to 10 is given. Initially, the algorithm does not know where the optimum is, which is represented by a uniform distribution (b) over the elements. After the execution of the algorithm, the solution was found, which is represented by a distribution (c) concentrating its probability mass on the maximum.

If $N$ is small, say $N = 10$, then choosing the largest number is easy, because one can simply revise the whole array and then pick the largest number. However, if $N$ is very large, say $N = 1000$, then comparing all the numbers becomes a difficult task. How do we go about this problem then? One solution would be to limit ourselves to comparing only a fraction of the array, say $M \ll N$ elements. This reduces the computational complexity to $\mathrm{O}(M)$ time and $\mathrm{O}(\log N)$ space at the cost of tolerating an error with probability $1 - M/N$. That is, we can give up certainty for the sake of computational efficiency. Notice that the time-space product is $\mathrm{O}(M \log N)$, which is linear in $M$ for fixed $N$.

To understand the effect of the resulting tradeoff, we again analyze how the knowledge state is transformed. We assume that the set was shuffled beforehand and that the algorithm inspects the first $M$ outcomes in linear order, choosing the largest number. We then use the frequency of choosing each number as its probability. This assigns equal probability to each one of the $N!$ possible permutations of the array. The resulting probability distributions are shown in Figure 7.2. Here, we see that inspecting only one element does not change the state of knowledge at all; that increasing $M$ moves the probability mass towards the larger numbers; and that complete certainty is achieved when $M = N$. Furthermore, notice that we can actually infer the ranking of the numbers by merely looking at the distributions, since larger numbers consistently get more probability mass.

The particular shape of the resulting distributions has some interesting properties. Let $\rho$ denote the relative entropy between the initial and the final distribution, and let $C = M - 1$ denote the number of comparisons carried out by the algorithm, and let $V$ denote the expected maximum. Note that $C = \mathrm{O}(M \log N)$ for fixed $N$, i.e. it serves as a measure of the computational complexity. If we plot $\rho$ versus $C$ (Figure 7.3a), we see

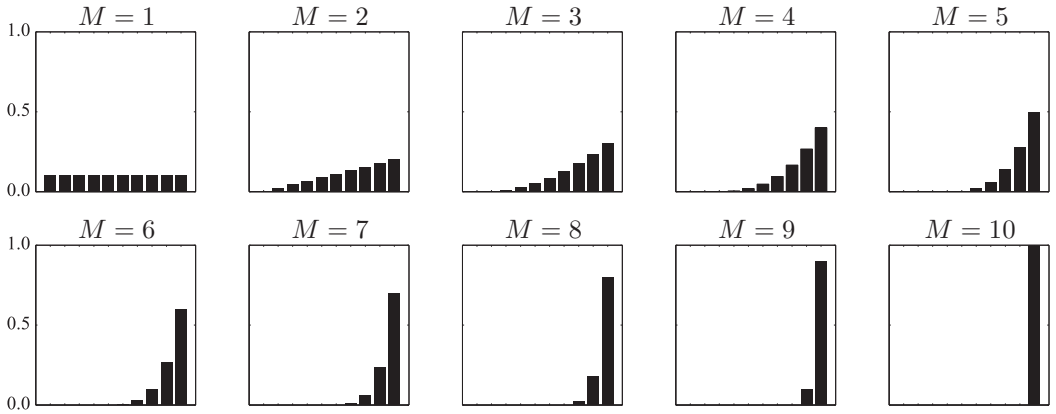



**Figure 7.2:** Distributions after Bounded Optimization. The plots show the distributions over the maximum in $\{1, 2, \ldots, 10\}$ obtained after running the bounded optimization algorithm for different values of $M$.

a surprising property: the quantities are proportional, having a correlation coefficient[1] of $r = 0.9991$. That is, we can use $\rho$ as a good measure of the computational complexity of the algorithm. We also want to understand how the expected value $V$ evolves as we increase the computational effort. This is seen by plotting $V$ versus $\rho$ (Figure 7.3b). This plot shows that certainty has a *diminishing marginal value*, that is, the gain in the value decreases with more computation. Intuitively, this is because the better the candidate solution, the more effort it takes to find an even better one. Moreover, the shape of the expected utility turns out to be logarithmic, as can be seen by plotting $2^V$ versus $\rho$.

Intuitively, if we care about computational costs, then it is not always a good idea to run an exhaustive optimization algorithm, because we can reach a point where the gain in value is too small to justify the extra computational effort. This idea can be captured by changing the evaluation criterion to one that penalizes the expected value by the relative entropy, e.g.

$$V - \alpha\rho,$$

where $\alpha$ is a conversion factor. Fixing $\alpha$ defines a tradeoff between the expected value and the relative entropy (Figure 7.4). The plot confirms our intuition: the smaller $\alpha$, the larger the number of elements we compare, and the more "rational" our choice becomes. A brief simulation also shows that a achieving a perfectly rational choice does not require $\alpha = 0$; it is already achieved for $\alpha \approx 0.2131$.

This analysis leads to a principled algorithm for bounded optimization. For a fixed $\alpha$, consider the algorithm that linearly inspects the elements of the array until

[1] More precisely, the *Pearson product-moment correlation coefficient* is an indicator of the linear dependence between two variables, having values ranging from $-1$ to 1. A value equal to 1 means that a linear relation perfectly describes the relationship between the two variables.

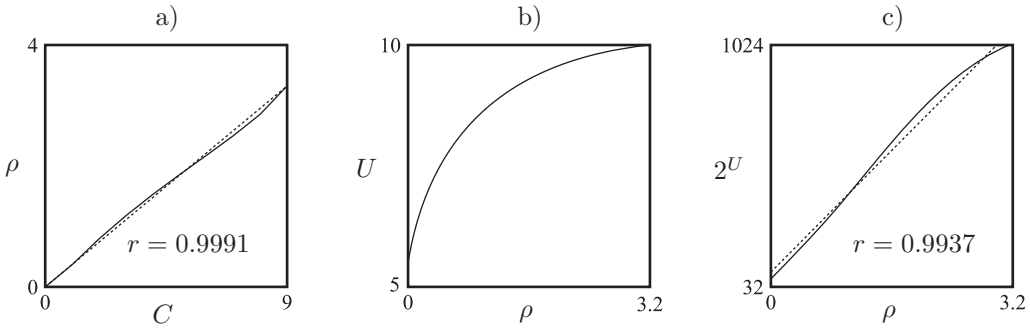



**Figure 7.3:** Performance of the Bounded Optimization. In panel (a), it is apparent that the relative entropy $\rho$ is proportional to the number of comparisons $C$ carried out by the algorithm, and hence $\rho$ serves as a measure of the computational complexity. Plot (b) shows the expected value $V$ against the relative entropy $\rho$. The marginal increment of $V$ diminishes as $\rho$ increases. Furthermore, panel (c) shows that $2^V$ is linear in $\rho$, meaning that $V$ is logarithmic in $\rho$.

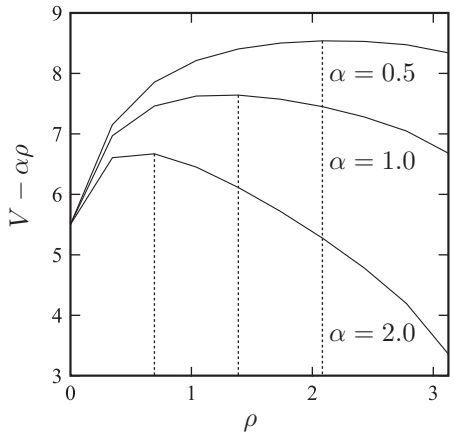



**Figure 7.4:** Expected Value Penalized by Relative Entropy. Choosing the conversion factor $\alpha$ defines a tradeoff between the expected value and the relative entropy. In the plot, three performance curves are shown, corresponding to the tradeoffs $\alpha = \frac{1}{2}, 1$ and 2. Notice how an exponential increase of $\alpha$ leads to a linear increase in the optimal $\rho$.

the largest number found so far, penalized by $\alpha\rho$, reaches its peak. This algorithm is stochastic, with the property that the probabilities of choosing a number are monotonic: for all $i, j$,

$$p_i > p_j \quad \Longleftrightarrow \quad v_i > v_j$$

where $v_i$ is the value of the $i$-th element and and $p_i$ is its probability of being chosen. This seems to be a more general property of a bounded optimization algorithm. Furthermore, notice that because of the diminishing marginal value, it is easy to reach a good performance level (although it is very hard to approach the optimum).

This concludes our example. In the remainder of this chapter, we aim to develop a general framework of bounded rationality and then apply it to autonomous systems. We are especially interested in providing a solid axiomatic basis for bounded rationality.

## 7.2 Utility & Resources

In physics, the behavior of a system can be described in two ways: using the dynamical equations or using an extremal principle (Goldstein, 1980). The first one specifies how the coordinates of the physical system change in time, like e.g. in Newton's second law

$$\mathbf{F} = \frac{d\mathbf{p}}{dt}$$

where $F$ is the force, $\mathbf{p}$ is the momentum and $t$ is time. The second expresses the dynamics of a system as the solution to a variational problem, like e.g. the action integral

$$A = \int_{t_i}^{t_f} L[\mathbf{q}, \dot{\mathbf{q}}, t]\, dt$$

in Langrangian Mechanics, where $L$ is the Langrangian (i.e. the kinetic energy minus the potential energy of the system), $\mathbf{q}$ are the generalized coordinates of the system and $t$ is time. According to the principle of least action, the dynamical equations of the system are then obtained by finding the trajectory $\mathbf{q}(t)$ that is an extremum of the action integral. On a conceptual level, the difference between stating the dynamical equations or the extremum principle to specify a physical system is analogous to the difference between stating the output model or the subjective expected utility in order to specify an autonomous system.

In the previous chapter, we have argued that the resource cost (i.e. work) of observing an event $A$ given $B$ is

$$-\gamma \log \mathbf{P}(A|B),$$

where $\gamma > 0$ is the conversion factor between units of energy and information. This has the advantage of linking three disparate concepts together, namely resource costs (physics), information content (information theory) and behavior $\mathbf{P}(A|B)$ (statistics). Can we exploit this connection to physics in order to devise a new principle for constructing autonomous systems? The answer is yes, but this connection is not straightforward, because it requires revising the way we think about utilities and about the process of constructing an autonomous system.

### 7.2.1 Utility

In this section, we propose a non-standard concept of utility that is more in accord with thermodynamics and information-theory. Let $\mathbf{U}(A)$ denote the utility of an event $A \in \mathcal{F}$ and let

$$\mathbf{u}(A|B) := \mathbf{U}(A \cap B) - \mathbf{U}(B)$$

be a shorthand to denote the gain in utility obtained from experiencing event $A$ given event $B$. We will derive the relation

$$\mathbf{u}(A|B) = \alpha \log \mathbf{P}(A|B)$$

where $\alpha > 0$ is a conversion factor between units of utility and information. This relation is obtained from desiderata characterizing the notion of utilities in systems that are "free" to generate events. More specifically, by a **free system** we mean a system that can choose the sampling probabilities of events generated by itself.

Consider a free system represented by a finite probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Here, the probability measure $\mathbf{P}$ models the generative law that the system uses to choose events. Thus, if $\mathbf{P}(A) > \mathbf{P}(B)$, then the propensity of experiencing $A$ is higher than $B$. In such a system, differences in probability can be given an interpretation relating them to differences in preferences: one can say that $A$ is more probable than $B$ because $A$ is more desirable than $B$. In other words, a system is more likely to choose the events that it prefers. Using this line of reasoning, can we find a quantity which will measure how desirable an event is? We call such a measure of a **utility gain function**, although one should always bear in mind that the resulting utilities do not correspond to the same notion of utility that we have seen in Part I.

If there is such a measure, then it is reasonable to demand the following three properties for it:

i. Utility gains should be mappings from conditional probabilities into real numbers.

ii. Utility gains should be additive. That is, the gain of a joint event should be obtained by summing up the gain of the sub-events. For example, the "gain of drinking coffee and eating a croissant" should equal "the gain of drinking coffee" plus the "gain of having a croissant given the gain of drinking coffee".

iii. A more probable event should have a higher utility gain than a less probable event. For example, if "drinking a coffee" is more likely than "eating a croissant", then the utility gain of the former is higher. Note that this is not necessarily true anymore if the system is not free to choose between events. For instance, "losing the lottery" has a much higher probability than "winning the lottery" even though the utility gain of the latter is obviously higher, but this case differs from the previous example in that the event is not determined by the system itself.

The three properties can then be summarized as follows.

**Definition 22 (Axioms of Utility)** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A set function $\mathbf{U} : \mathcal{F} \to \mathbb{R}$ is a **utility function** for $\mathbf{P}$ iff its **utility gain function** $\mathbf{u}(A|B) := \mathbf{U}(A \cap B) - \mathbf{U}(B)$ has the following three properties for all events $A, B, C, D \in \mathcal{F}$:

| | | |
|---|---|---|
| U1. | $\exists f, \mathbf{u}(A\|B) = f\big(\mathbf{P}(A\|B)\big) \in \mathbb{R},$ | (real-valued) |
| U2. | $\mathbf{u}(A \cap B\|C) = \mathbf{u}(A\|C) + \mathbf{u}(B\|A \cap C),$ | (additive) |
| U3. | $\mathbf{P}(A\|B) > \mathbf{P}(C\|D) \quad \Leftrightarrow \quad \mathbf{u}(A\|B) > \mathbf{u}(C\|D).$ | (monotonic) |

Furthermore, we use the abbreviation $\mathbf{u}(A) := \mathbf{u}(A|\Omega)$. □

The following theorem shows that these three properties enforce a strict mapping between probabilities and utility gains.

**Theorem 5 (Utility Gain ↔ Probability)** *If $f$ is such that $\mathbf{u}(A|B) = f(\mathbf{P}(A|B))$ for every probability space $(\Omega, \mathcal{F}, \mathbf{P})$, then $f$ is of the form*

$$f(\cdot) = \alpha \log(\cdot),$$

*where $\alpha > 0$ is arbitrary strictly positive constant.* □

Proof Given arbitrary $p, q \in (0, 1]$ and $n, m \in \mathbb{N}$, one can always choose a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ having sequences of events $A_1, A_2, \ldots, A_n \in \mathcal{F}$ and $B_1, B_2, \ldots, B_m \in \mathcal{F}$, and an event $C \in \mathcal{F}$ such that

$$\begin{aligned} p &= \mathbf{P}(A_1|C) = \mathbf{P}(A_2|C \cap A_1) = \cdots = \mathbf{P}(A_n|C \cap A_{<n}), \\ q &= \mathbf{P}(B_1|C) = \mathbf{P}(B_2|C \cap B_1) = \cdots = \mathbf{P}(B_m|C \cap B_{<m}), \end{aligned}$$

where we use the shorthand $A_{<j} := \bigcap_{i<j} A_i$. Applying $f$ to the first sequence yields the equivalence

$$\mathbf{P}(A_1|C) = \mathbf{P}(A_i|C \cap A_{<j}) \qquad \Longleftrightarrow \qquad \mathbf{u}(A_1|C) = \mathbf{u}(A_i|C \cap A_{<j}) \tag{7.1}$$

for all $i = 1, \ldots, n$. Then, one can show

$$\begin{aligned} f\big(\mathbf{P}(A_1|C)^n\big) &\overset{(a)}{=} f\left(\prod_{i=1}^{n} \mathbf{P}(A_i\Big|C \cap A_{<j})\right) \overset{(b)}{=} f\big(\mathbf{P}(A_1 \cap \cdots \cap A_n|C)\big) \\ &\overset{(c)}{=} \mathbf{u}(A_1 \cap \cdots A_n|C) \overset{(d)}{=} \sum_{i=1}^{n} \mathbf{u}(A_i|C \cap A_{<j}) \\ &\overset{(e)}{=} n\mathbf{u}(A_1|C) \overset{(f)}{=} nf(\mathbf{P}(A_1|C)). \end{aligned}$$

Equality (a) is obtained by substituting each $\mathbf{P}(A_1|C)$ in the product $\mathbf{P}(A_1|C)^n$ with a corresponding $\mathbf{P}(A_i|C \cap A_{<j})$ term. Equality (b) is obtained by repeatedly applying the product rule. Applying first the function to transform the probability into a utility gain and then using the additivity property yields equalities (c) and (d). Finally,

equalities (e) and (f) are obtained by first using (7.1) and then transforming back to probabilities. Similarly, for the second sequence, one has

$$f\big(\mathbf{P}(B_1|C)^m\big) = mf(\mathbf{P}(B_1|C)).$$

Since $p, q$ and $n, m$ are arbitrary, then there is always a probability space such that

$$f(p^n) = nf(p) \quad \text{and} \quad f(q^m) = mf(q).$$

This part of the argument parallels Shannon's entropy theorem (Shannon, 1948). Let $p, q \in (0, 1]$ such that $q < p$. Choose an arbitrarily large $m \in \mathbb{N}$ and find an $n \in \mathbb{N}$ to satisfy $q^m < p^n < q^{m+1}$. Taking the logarithm, and dividing by $n \log q$ one obtains

$$\frac{m}{n} < \frac{\log p}{\log q} < \frac{m}{n} + \frac{1}{n}. \tag{7.2}$$

Similarly, using $f(p^n) = nf(p)$ and the monotonicity of $f$, we have

$$\begin{array}{rccccc} & q^m & < & p^n & < & q^{m+1} \\ \iff & f(q^m) & < & f(p^n) & < & f(q^{m+1}) \\ \iff & mf(q) & < & nf(p) & < & (m+1)f(q). \end{array}$$

Dividing the last set of inequalities by $nf(p)$ yields

$$\frac{m}{n} < \frac{f(p)}{f(q)} < \frac{m}{n} + \frac{1}{n}. \tag{7.3}$$

Combining the inequalities in (7.2) and (7.3), one gets

$$\Big|\frac{\log p}{\log q} - \frac{f(p)}{f(q)}\Big| < \frac{2}{n}.$$

Since $m, n$ can be chosen arbitrary large, this implies

$$\frac{\log p}{\log q} = \frac{f(p)}{f(q)}$$

in the limit $n \to \infty$. Fixing $q$ and rearranging terms gives the functional form

$$f(p) = \alpha \log p,$$

where $\alpha$ must be positive to satisfy the monotonicity property.

The previous choice of probability spaces implies that $f(p) = \alpha \log p$. We show now that this choice does not violate the axioms for any probability space. First, $f$ is real valued. Second,

$$\begin{aligned} \mathbf{u}(A \cap B|C) &= \alpha \log \mathbf{P}(A \cap B|C) = \alpha \log\big(\mathbf{P}(A|C)\mathbf{P}(B|A \cap C)\big) \\ &= \alpha \log \mathbf{P}(A|C) + \alpha \log \mathbf{P}(B|A \cap C) = \mathbf{u}(A|C) + \mathbf{u}(B|A \cap C), \end{aligned}$$

where the equalities are obtained by using the product rule for probabilities and a property of logarithms. Third, $f$ is monotonic. Hence, $f(p) = \alpha \log p$ holds for any probability space. ∎

Thus, if one is willing to accept Definition 22 as a reasonable characterization of a utility function, then one obtains the relations

$$\mathbf{U}(A \cap B) - \mathbf{U}(B) = \alpha \log \mathbf{P}(A|B). \tag{7.4}$$

In this relation, $\alpha > 0$ plays the role of a conversion factor between utilities and information. If a probability measure $\mathbf{P}$ and a utility function $\mathbf{U}$ satisfy the relation (7.4), then we say that they are **conjugate**. We call one unit of utility one **utile**. Given that this transformation between utility gains and probabilities is a bijection, one has that

$$\mathbf{P}(A|B) = 2^{\frac{1}{\alpha}(\mathbf{U}(A \cap B) - \mathbf{U}(B))}.$$

There are several important observations with respect to this particular functional form.

First, recall that in Bayesian probability theory (Section 4.1), a new measurement $A$ is combined with the previous knowledge $B$ by an intersection, i.e. the posterior knowledge is given by $A \cap B$. Furthermore, note that utility gains are always negative, i.e.

$$\mathbf{U}(A \cap B) - \mathbf{U}(B) \leq 0.$$

This means that every measurement decreases the utility. While this sounds counterintuitive, the relation also says that minimizing the resource costs maximizes the utility that is achieved after the interaction. Hence, if the free system is forced to act, then it will favor outcomes that reduce the utility less.

Second, note that in the context of thermodynamics discussed in Section 6.2.1, doing work $W = -\gamma \log \mathbf{P}(A|B)$ on a physical system changes its energy level from $\mathbf{e}(B)$ to $\mathbf{e}(A \cap B)$ as

$$\mathbf{e}(B) \quad \longrightarrow \quad \mathbf{e}(A \cap B) = \mathbf{e}(B) + W = \mathbf{e}(B) - \gamma \log \mathbf{P}(A|B).$$

That is, the work done on a system increases its internal energy. Comparing this with the relation

$$\mathbf{U}(A \cap B) - \mathbf{U}(B) = -\alpha \log \mathbf{P}(A|B)$$

leads to the conclusion that

$$\mathbf{U}(A) = -\tfrac{\alpha}{\gamma}\mathbf{e}(A)$$

for all $A \in \mathcal{F}$. Hence, the utilities that we have derived from mathematical desiderata are (safe for a conversion factor) negative energy levels. This is useful because it allows us analyzing decision making by borrowing ideas from thermodynamics.

Furthermore, note that the exponentiation of the utility function can be written as a sum of parts. That is, if $A_1, A_2 \in \mathcal{F}$ form a partition of $A \in \mathcal{F}$, then

$$2^{\frac{1}{\alpha}\mathbf{U}(A)} = 2^{\frac{1}{\alpha}\mathbf{U}(A_1)} + 2^{\frac{1}{\alpha}\mathbf{U}(A_2)}.$$

This is because

$$\mathbf{P}(A|\Omega) = \mathbf{P}(A_1|\Omega) + \mathbf{P}(A_2|\Omega),$$

$$\frac{2^{\frac{1}{\alpha}\mathbf{U}(A)}}{2^{\frac{1}{\alpha}\mathbf{U}(\Omega)}} = \frac{2^{\frac{1}{\alpha}\mathbf{U}(A_1)}}{2^{\frac{1}{\alpha}\mathbf{U}(\Omega)}} + \frac{2^{\frac{1}{\alpha}\mathbf{U}(A_2)}}{2^{\frac{1}{\alpha}\mathbf{U}(\Omega)}},$$

$$2^{\frac{1}{\alpha}\mathbf{U}(A)} = 2^{\frac{1}{\alpha}\mathbf{U}(A_1)} + 2^{\frac{1}{\alpha}\mathbf{U}(A_2)}.$$

In particular, one can rewrite any probability $\mathbf{P}(A|B)$ as a Gibbs measure:

$$\mathbf{P}(A|B) = \frac{\sum_{\omega\in A} 2^{\frac{1}{\alpha}\mathbf{U}(\omega)}}{\sum_{\omega\in B} 2^{\frac{1}{\alpha}\mathbf{U}(\omega)}}.$$

where we have used the abbreviation $\mathbf{U}(\omega) := \mathbf{U}(\{\omega\})$. As the conversion factor $\alpha$ approaches zero, the probability measure $\mathbf{P}(\omega)$ approaches a uniform distribution over the maximal set $\Omega_{\max} := \{\omega^* \in \Omega | \omega^* = \arg\max_\omega \mathbf{U}(\omega)\}$. Similarly, as $\alpha \to \infty$, $\mathbf{P}(\omega) \to \frac{1}{|\Omega|}$, i.e. the uniform distribution over the whole outcome set $\Omega$. Also, note that

$$2^{\frac{1}{\alpha}\mathbf{U}(\Omega)} = \sum_\omega 2^{\frac{1}{\alpha}\mathbf{U}(\omega)}.$$

Intuitively, the utility $\mathbf{U}(\Omega)$ of the sample set corresponds to the utility of the system before any interaction has occurred.

### 7.2.2 Variational principle

The conversion between probability and utility established in the previous section can be characterized using a variational principle. Inspired by thermodynamics, we now define a quantity that will allow us analyzing changes of the state of the system.

**Definition 23 (Free Utility)** Let $(\Omega, \mathcal{F})$ be a measurable space and let $\mathbf{U}$ be a utility function. Define the **free utility** functional as

$$\mathbf{J}(\mathbf{Pr}; \mathbf{U}) := \sum_{\omega\in\Omega} \mathbf{Pr}(\omega)\mathbf{U}(\omega) - \alpha \sum_{\omega\in\Omega} \mathbf{Pr}(\omega) \log \mathbf{Pr}(\omega),$$

(the term $\mathbf{Pr}(\omega) := \mathbf{Pr}(\{\omega\})$ is an abbreviation) where $\mathbf{Pr}$ is an arbitrary probability measure over $(\Omega, \mathcal{F})$. □

Here, we see that the free utility[2] is the expected utility of the system plus the uncertainty (i.e. the entropy) over the outcome multiplied by the utility-information conversion factor. It satisfies the following relation.

[2]The functional $\mathbf{F} := -\mathbf{J}$ is also known as the *Helmholtz free energy* in thermodynamics. $\mathbf{F}$ is a measure of the "useful" work obtainable from a closed thermodynamic system at a constant temperature and volume.

**Theorem 6 (Variational Principle)** *A conjugate pair* $(\mathbf{P}, \mathbf{U})$ *satisfies*

$$\mathbf{J}(\mathbf{Pr}; \mathbf{U}) \leq \mathbf{J}(\mathbf{P}; \mathbf{U}) = \mathbf{U}(\Omega).$$

*where* $\mathbf{Pr}$ *is an arbitrary probability measure over* $\Omega$. □

PROOF Rewriting terms using the utility-probability conversion and applying Jensen's inequality yields

$$\begin{aligned}
\mathbf{J}(\mathbf{Pr}; \mathbf{U}) &= \sum_{\omega \in \Omega} \mathbf{Pr}(\omega)\mathbf{U}(\omega) - \alpha \sum_{\omega \in \Omega} \mathbf{Pr}(\omega) \log \mathbf{Pr}(\omega) \\
&= \alpha \sum_{\omega \in \Omega} \mathbf{Pr}(\omega) \log \frac{2^{\frac{1}{\alpha}\mathbf{U}(\omega)}}{\mathbf{Pr}(\omega)} \\
&\leq \alpha \log \sum_{\omega \in \Omega} \mathbf{Pr}(\omega) \frac{2^{\frac{1}{\alpha}\mathbf{U}(\omega)}}{\mathbf{Pr}(\omega)} \\
&= \alpha \log \sum_{\omega \in \Omega} 2^{\frac{1}{\alpha}\mathbf{U}(\omega)} \\
&= \mathbf{U}(\Omega),
\end{aligned}$$

with equality iff $\frac{2^{\frac{1}{\alpha}\mathbf{U}(\omega)}}{\mathbf{Pr}(\omega)}$ is constant, i.e. if $\mathbf{Pr} = \mathbf{P}$. ■

The variational principle tells us that the probability law $\mathbf{P}$ of the system is the one that maximizes the free utility for a given utility function $\mathbf{U}$, since

$$\mathbf{P} = \arg\max_{\mathbf{Pr}} \mathbf{J}(\mathbf{Pr}; \mathbf{U}).$$

Hence, the utility function $\mathbf{U}$ plays the role of a constraint landscape for probability measures $\mathbf{Pr}$ out of which the conjugate probability measure $\mathbf{P}$ is the one that maximizes the uncertainty.

### 7.2.3 Bounded SEU

In SEU theory, the designer constructs an autonomous system by first assuming a probability measure that characterizes the behavior of the environment and a utility criterion to compare different outcomes, and then by calculating a policy that maximizes the subjective expected utility. During this process, the calculation of the optimal policy is done disregarding the costs of the computation. As we have argued at the beginning of this chapter, here we do want to take into account the costs of computing the optimal policy. However, this poses two important problems.

First, how do we measure the costs of this computation? To answer this question, we first have to agree on what this calculation is actually doing. We know that after this calculation, the agent ends up with an optimal policy. But what if this calculation

had been omitted? Clearly, the agent would not end up with an optimal policy—unless the agent already had the optimal policy from the very beginning. The important conclusion here is that, in the general case, the purpose of such a calculation is to transform an initial system $\mathbf{P}_i$ into a final system $\mathbf{P}_f$. In the previous chapter, we have seen that the amount of bits that have to be set to carry out this transformation is given by the relative entropy, i.e.

$$\sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{P}_i(\omega)}.$$

Hence, the cost measured in joules and in utiles is obtained by multiplying the previous quantity by $\gamma$ and $\alpha$ respectively.

Second, how do we optimally choose this transformation? In other words, how should we transform the initial system $\mathbf{P}_i$ such that the final system $\mathbf{P}_f$ optimally trades off the benefits of maximizing a given utility function $\mathbf{U}_*$ against the cost of the transformation? To answer this question, consider the transformation represented in Figure 7.5. The initial system satisfies the equation

$$\mathbf{J}_i := \sum_{\omega\in\Omega} \mathbf{P}_i(\omega)\mathbf{U}_i(\omega) - \alpha \sum_{\omega\in\Omega} \mathbf{P}_i(\omega) \log \mathbf{P}_i(\omega) = \mathbf{U}_i(\Omega).$$

We add new constraints represented by the utility function $\mathbf{U}_*$. Then, the resulting utility function $\mathbf{U}_f$ is given by the sum

$$\mathbf{U}_f = \mathbf{U}_i + \mathbf{U}_*.$$

The free utility $\mathbf{J}_f$ of the final system is then given by

$$\begin{aligned}
\mathbf{J}_f &:= \sum_{\omega\in\Omega} \mathbf{P}_f(\omega)\mathbf{U}_f(\omega) - \alpha \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \mathbf{P}_f(\omega) \\
&= \sum_{\omega\in\Omega} \mathbf{P}_f(\omega)(\mathbf{U}_i(\omega) + \mathbf{U}_*(\omega)) - \alpha \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \mathbf{P}_f(\omega) \\
&= \sum_{\omega\in\Omega} \mathbf{P}_f(\omega)\mathbf{U}_*(\omega) - \alpha \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{P}_i(\omega)} + \mathbf{U}_i(\Omega).
\end{aligned}$$

To understand the change that has occurred due to this transformation, we take the difference $\mathbf{J}_f - \mathbf{J}_i$ in the free utility. This results in a quantity that will play a central role in our next development.

**Definition 24 (Bounded Subjective Expected Utility)** Let $\mathbf{U}_*$ a utility function and let $\mathbf{J}_i$ and $\mathbf{J}_f$ be the free utility for the conjugate pairs $(\mathbf{P}_i, \mathbf{U}_i)$ and $(\mathbf{P}_f, \mathbf{U}_f)$. Then, the **bounded subjective expected utility** (bounded SEU) is given by the difference in free utility

$$\mathbf{J}_f - \mathbf{J}_i = \sum_{\omega\in\Omega} \mathbf{P}_f(\omega)\mathbf{U}_*(\omega) - \alpha \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{P}_i(\omega)}. \tag{7.5}$$

□

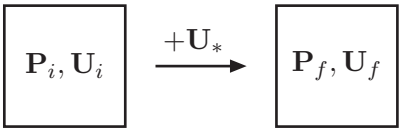



**Figure 7.5:** Transformation of a System. A transformation from a system $(\mathbf{P}_i, \mathbf{U}_i)$ into a system $(\mathbf{P}_f, \mathbf{U}_f)$ by addition of a constraint $\mathbf{U}_*$.

This difference in free utility has an interpretation that is crucial for the formalization of bounded rationality: it is the expected target utility $\mathbf{U}_*$ (first term) penalized by the cost of transforming $\mathbf{P}_i$ into $\mathbf{P}_f$ (second term). In practice, this means the following. A designer starts out with an initial system $\mathbf{P}_i$ that he wants to change in order to maximize a utility function $\mathbf{U}_*$. He then changes this system into $\mathbf{P}_f$, spending in the process a total amount of

$$\gamma \sum_{\omega \in \Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{P}_i(\omega)}$$

joules of energy. The expected utility of the system is given by

$$\sum_{\omega \in \Omega} \mathbf{P}_f(\omega) \mathbf{U}_*(x).$$

However, from the point of view of the designer, the total expected utility is smaller, because he has to subtract the reduction in utility caused by the cost of the transformation itself:

$$\sum_{\omega \in \Omega} \mathbf{P}_f(\omega) \mathbf{U}_*(\omega) - \alpha \sum_{\omega \in \Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{P}_i(\omega)}.$$

**Remark 20** Alternatively, one can interpret bounded SEU as the expected utility penalized by an "uncertainty" term. In this interpretation, the relative entropy measures the "risk" of a gamble. □

Because of its interpretation, we define (7.5) as a functional to be maximized, either with respect to $\mathbf{P}_f$ or with respect to $\mathbf{P}_i$. We call this construction method the **maximum bounded SEU principle**. Depending on whether we vary $\mathbf{P}_f$ or $\mathbf{P}_i$, one obtains two different variational problems having different applications.

**Control Method.** The construction method **for control** corresponds to the case when the designer wants to build a system that optimizes a given utility function $\mathbf{U}_*$ subject to the costs of the transformation. This is achieved by fixing $\mathbf{P}_i$ and varying $\mathbf{P}_f$:

$$\mathbf{P}_f = \arg\max_{\mathbf{Pr}} \sum_{\omega \in \Omega} \mathbf{Pr}(\omega) \mathbf{U}_*(\omega) - \alpha \sum_{\omega \in \Omega} \mathbf{Pr}(\omega) \log \frac{\mathbf{Pr}(\omega)}{\mathbf{P}_i(\omega)}. \tag{7.6}$$

The solution is given by

$$\mathbf{P}_f(\omega) \propto \mathbf{P}_i(\omega) \exp\left(\frac{1}{\alpha}\mathbf{U}_*(\omega)\right).$$

This can be interpreted as follows. The decision maker starts out with a prior probability of choosing $\omega$ given by $\mathbf{P}_i(\omega)$. Then, he changes this probability to $\mathbf{P}_f(\omega)$ obtained from multiplying the prior with the term $\exp(\frac{1}{\alpha}\mathbf{U}_*(\omega))$, which intuitively corresponds to the "likelihood" of $\omega$ being the best choice. Compare this to the example given in the introduction. Here, $\alpha$ controls the amount of resource units required to increase the utility. In particular, when the conversion factor between information and utility is negligible $\alpha \approx 0$, then (7.5) becomes

$$\mathbf{J}_f - \mathbf{J}_i \approx \sum_{\omega\in\Omega} \mathbf{P}_f(\omega)\mathbf{U}_*(x),$$

and hence resource costs are ignored in the choice of $\mathbf{P}_f$, leading to $\mathbf{P}_f \approx \delta_{\omega^*}(\omega)$, where $\omega^* = \arg\max_\omega \mathbf{U}_*(\omega)$. Similarly, when $\alpha$ is very high, then the difference is

$$\mathbf{J}_f - \mathbf{J}_i \approx -\alpha \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{P}_i(\omega)},$$

and hence no computation is carried out to optimize the choice, i.e. $\mathbf{P}_f \approx \mathbf{P}_i$.

**Estimation Method.** The construction method **for estimation** corresponds to the case when the designer wants to build a system that approximates another system (that is maximizing a possibly unknown utility function $\mathbf{U}_*$) subject to the costs of the transformation. This is achieved by fixing $\mathbf{P}_f$ and varying $\mathbf{P}_i$:

$$\begin{aligned}\mathbf{P}_i &= \arg\max_{\mathbf{Pr}} \sum_{\omega\in\Omega} \mathbf{P}_f(\omega)\mathbf{U}_*(\omega) - \alpha \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{Pr}(\omega)} \\ &= \arg\min_{\mathbf{Pr}} \sum_{\omega\in\Omega} \mathbf{P}_f(\omega) \log \frac{\mathbf{P}_f(\omega)}{\mathbf{Pr}(\omega)},\end{aligned} \tag{7.7}$$

and thus we have recovered the minimum relative entropy principle for estimation, having the solution

$$\mathbf{P}_i = \mathbf{P}_f$$

and therefore carry no transformation costs. While this "construction method" might look bizarre at a first glance, it is saying something obvious: If the designer is looking for a system $\mathbf{P}_i$ that approximates another known system $\mathbf{P}_f$, then the best he can do is to just choose $\mathbf{P}_f$ itself without having to carry out any transformation at all!

## 7.3 Bounded SEU in Autonomous Systems

We tackle now the question of how to construct an autonomous system using the bounded SEU principle. In the previous section, we have learnt that building a system is conceptualized as transforming a preexisting system. Furthermore, there are two construction methods, namely one for control and one for estimation.

We assume that we are in possession of a reference I/O model $\mathbf{P}_0$ and a utility function $\mathbf{U}_*$. The objective is to find an I/O model $\mathbf{P}$ maximizing the bounded SEU using the method for control, estimation or a mixture of both. We further assume that the conversion factor between utilities and information is given by $\alpha > 0$.

To construct the associated bounded SEU, one has to carefully decide what construction method to apply for each random variable. For instance, consider constructing an I/O model $\mathbf{P}(ao)$ from a reference I/O model $\mathbf{P}_0(ao)$ using a utility function $\mathbf{U}_*$. Then, assuming that $a$ is derived using the control method and $o$ using the estimation method, one gets

$$\mathbf{P} = \arg\max_{\mathbf{Pr}} \left\{ \sum_{ao} \mathbf{Pr}(a)\mathbf{P}_0(o|a)\mathbf{U}_*(ao) - \alpha \sum_{ao} \mathbf{Pr}(a)\mathbf{P}_0(o|a) \log \frac{\mathbf{Pr}(a)\mathbf{P}_0(o|a)}{\mathbf{P}_0(a)\mathbf{Pr}(o|a)} \right\}.$$

This is because for $a$, the reference probabilities $\mathbf{P}_0(a)$ play the role of $\mathbf{P}_f$ in (7.5); and for $o$, the reference probabilities $\mathbf{P}_0(o|a)$ play the role of $\mathbf{P}_i$ in (7.5).

A simple way to concisely write down the bounded SEU for mixed methods is by defining two auxiliary I/O models $\mathbf{R}$ and $\mathbf{S}$ as

$$\mathbf{R}(a_t|\underline{ao}_{<t}) := \begin{cases} \mathbf{Pr}(a_t|\underline{ao}_{<t}) \\ \mathbf{P}_0(a_t|\underline{ao}_{<t}) \end{cases} \qquad \mathbf{R}(o_t|\underline{ao}_{<t}a_t) := \begin{cases} \mathbf{Pr}(o_t|\underline{ao}_{<t}a_t) & \text{(control)} \\ \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) & \text{(estimation)} \end{cases}$$

$$\mathbf{S}(a_t|\underline{ao}_{<t}) := \begin{cases} \mathbf{P}_0(a_t|\underline{ao}_{<t}) \\ \mathbf{Pr}(a_t|\underline{ao}_{<t}) \end{cases} \qquad \mathbf{S}(o_t|\underline{ao}_{<t}a_t) := \begin{cases} \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) & \text{(control)} \\ \mathbf{Pr}(o_t|\underline{ao}_{<t}a_t) & \text{(estimation)} \end{cases}$$

Then, the bounded SEU is given by

$$\mathbf{P} = \arg\max_{\mathbf{Pr}} \left\{ \sum_{\underline{ao}_{\leq T}} \mathbf{R}(\underline{ao}_{\leq T})\mathbf{U}_*(\underline{ao}_{\leq T}) - \alpha \sum_{\underline{ao}_{\leq T}} \mathbf{R}(\underline{ao}_{\leq T}) \log \frac{\mathbf{R}(\underline{ao}_{\leq T})}{\mathbf{S}(\underline{ao}_{\leq T})} \right\}. \tag{7.8}$$

Observe that another way of looking at (7.8) is as a collection of independent variational problems, where this collection contains one variational problem for each $\mathbf{P}(a_t|\underline{ao}_{<t})$ and $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$. We now show two examples of how to construct autonomous systems using the bounded SEU principle.

### 7.3.1 Bounded Optimal Control

As we have seen in Section 3.2.4, in optimal control problems it is generally assumed that we are given a utility function $\mathbf{U}_*$ and that the environment is fully known, i.e.

$\mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) = \mathbf{Q}(o_t|\underline{ao}_{<t}a_t)$. The probability measures $\mathbf{R}$ and $\mathbf{S}$ are given by

$$\begin{aligned} \mathbf{R}(a_t|\underline{ao}_{<t}) &= \mathbf{Pr}(a_t|\underline{ao}_{<t}), & \mathbf{R}(o_t|\underline{ao}_{<t}a_t) &= \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t), \\ \mathbf{S}(a_t|\underline{ao}_{<t}) &= \mathbf{P}_0(a_t|\underline{ao}_{<t}), & \mathbf{S}(o_t|\underline{ao}_{<t}a_t) &= \mathbf{Pr}(o_t|\underline{ao}_{<t}a_t). \end{aligned}$$

Hence, the variational problem to find $\mathbf{P}$ is to maximize the functional

$$\sum_{\underline{ao}_{\leq T}} \mathbf{R}(\underline{ao}_{\leq T}) \left[ \mathbf{U}_*(\underline{ao}_{\leq T}) - \alpha \sum_{t=1}^{T} \log \frac{\mathbf{Pr}(a_t|\underline{ao}_{<t})}{\mathbf{P}_0(a_t|\underline{ao}_{<t})} - \alpha \sum_{t=1}^{T} \log \frac{\mathbf{P}_0(o_t|\underline{ao}_{<t}a_t)}{\mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)} \right], \quad (7.9)$$

which results from using the definition of $\mathbf{R}$ and $\mathbf{S}$ in (7.8). In the variational problem for the observation probabilities we can disregard the constraint utilities and the resource cost of the action probabilities. The $t$-th summand of the total expected reward can then be written as

$$\sum_{\underline{ao}_{<t}a_t} \mathbf{R}(\underline{ao}_{<t}a_t) \left[ \sum_{o_t} \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) \log \frac{\mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)}{\mathbf{P}_0(o_t|\underline{ao}_{<t}a_t)} \right].$$

Since varying $\mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)$ does not influence the summands at times $\neq t$, the optimal solution to this minimum relative entropy problem is trivially obtained by

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) = \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t).$$

The variational problem with respect to the action probabilities is a bit more intricate, since varying the action probability at time $t$ has an impact on all subsequent conditional action probabilities. The functional (7.9) can be expanded recursively, yielding

$$\begin{aligned} &\sum_{a_1} \mathbf{Pr}(a_1) \Bigg[ - \alpha \log \frac{\mathbf{Pr}(a_1)}{\mathbf{P}_0(a_1)} + \sum_{o_1} \mathbf{P}(o_1|a_1) \Bigg[ \\ &+ \sum_{a_2} \mathbf{Pr}(a_2|\underline{ao}_1) \Bigg[ - \alpha \log \frac{\mathbf{Pr}(a_2|\underline{ao}_1)}{\mathbf{P}_0(a_2|\underline{ao}_1)} + \sum_{o_2} \mathbf{P}(o_2|\underline{ao}_1 a_2) \Bigg[ \\ &+ \cdots \\ &+ \sum_{a_T} \mathbf{Pr}(a_T|\underline{ao}_{<T}) \Bigg[ - \alpha \log \frac{\mathbf{Pr}(a_T|\underline{ao}_{<T})}{\mathbf{P}_0(a_T|\underline{ao}_{<T})} + \sum_{o_T} \mathbf{P}(o_T|\underline{ao}_{<T}a_T) \mathbf{U}_*(\underline{ao}_{\leq T}) \Bigg] \cdots \Bigg] \Bigg] \Bigg] \Bigg], \end{aligned} \quad (7.10)$$

By inspection of the recursive expansion (7.10), one sees that the action probabilities $\mathbf{Pr}(a_t|\underline{ao}_{<t})$ can be solved backwards in time (akin to the Bellman optimality equations, Section 3.2.3). The innermost variational problem can be written as

$$\sum_{a_T} \mathbf{Pr}(a_T|\underline{ao}_{<T}) \left[ \sum_{o_T} \mathbf{P}(o_T|\underline{ao}_{<T}a_T) \mathbf{U}_*(\underline{ao}_{\leq T}) - \alpha \log \frac{\mathbf{Pr}(a_T|\underline{ao}_{<T})}{\mathbf{P}_0(a_T|\underline{ao}_{<T})} \right].$$

As discussed in Section 7.2.3, its solution is

$$\mathbf{P}(a_T|\underline{ao}_{<T}) = \frac{\mathbf{P}_0(a_T|\underline{ao}_{<T})}{Z^\alpha(\underline{ao}_{<T})} \exp\left\{\tfrac{1}{\alpha}\sum_{o_T}\mathbf{P}(o_T|\underline{ao}_{<T}a_T)\mathbf{U}_*(\underline{ao}_{\leq T})\right\},$$

where $Z^\alpha(\underline{ao}_{<T})$ is the normalizing constant, also known as the partition function. By induction, it is seen that the variational problem for the time steps $t < T$ can be written as

$$\sum_{a_t}\mathbf{Pr}(a_t|\underline{ao}_{<t})\left[\alpha\sum_{o_t}\mathbf{P}(o_t|\underline{ao}_{<t}a_t)\log Z^\alpha(\underline{ao}_{\leq t}) - \alpha\log\frac{\mathbf{Pr}(a_t|\underline{ao}_{<t})}{\mathbf{P}_0(a_t|\underline{ao}_{<t})}\right],$$

where the $Z^\alpha(\underline{ao}_{\leq t})$ are the normalizing constants obtained from subsequent time steps. The action probabilities $\mathbf{P}(a_t|\underline{ao}_{<t})$ for the times $t < T$ are given as

$$\mathbf{P}(a_t|\underline{ao}_{<t}) = \frac{\mathbf{P}_0(a_t|\underline{ao}_{<t})}{Z^\alpha(\underline{ao}_{<t})} \exp\left\{\sum_{o_t}\mathbf{P}(o_t|\underline{ao}_{<t}a_t)\log Z^\alpha(\underline{ao}_{\leq t})\right\}.$$

This way the optimal action probabilities can be computed recursively.

**Case $\alpha \to 0$.** We first investigate the limit case for $\alpha \to 0$. Identify the future utility $F^\alpha(\underline{ao}_{\leq t})$ using the recursion

$$\begin{aligned} F^\alpha(\underline{ao}_{\leq T}) &:= \mathbf{U}_*(\underline{ao}_{\leq T}), \\ F^\alpha(\underline{ao}_{\leq t}) &:= \alpha\log Z^\alpha(\underline{ao}_{\leq t}). \end{aligned}$$

If one takes the limit $\alpha \to 0$, then the future utility $F^\alpha(\underline{ao}_{<t})$ converges to the recursive formula

$$\begin{aligned} F^0(\underline{ao}_{\leq T}) &= \mathbf{U}_*(\underline{ao}_{\leq T}), \\ F^0(\underline{ao}_{<t}) &= \max_{a_t}\sum_{o_t}\mathbf{P}(o_t|\underline{ao}_{<t}a_t)F^0(\underline{ao}_{\leq t}), \end{aligned}$$

and the policy $\mathbf{P}(a_t|\underline{ao}_{<t})$ to[3]

$$\mathbf{P}(a_t|\underline{ao}_{<t}) = \delta^{a_t}_{a^*}, \quad \text{where} \quad a^* := \max_{a_t}\sum_{o_t}\mathbf{P}(o_t|\underline{ao}_{<t}a_t)F^0(\underline{ao}_{\leq t}).$$

Compare this to Definition 12. We have recovered the Bellman optimality equations for utilities! Using an analogous construction one can also recover the Bellman optimality equations for rewards.

**Case $\alpha \to \infty$.** Now we investigate the limit of unbounded resource costs by taking the limit $\alpha \to \infty$. This case yields

$$\mathbf{P}(a_t|\underline{ao}_{<t}) = \mathbf{P}_0(a_t|\underline{ao}_{<t}).$$

Hence, the optimal agent is the reference agent itself, as expected.

[3]This result holds assuming that in each step there is only one optimal action.

### 7.3.2 Adaptive Estimation

We consider the problem of adaptive estimation of an unknown source (i.e. environment). Let $P_0$ denote in the following a Bayesian input model with no actions, i.e. characterized by $P_0(\theta)$ and $P_0(o_t|\theta, o_{<t})$. This models possible input sources $P_0(o_t|\theta, o_{<t}) = \mathbf{Q}(o_t|\theta, o_{<t})$ indexed by $\theta \in \Theta$ and chosen randomly with probabilities $P_0(\theta)$. Let $\mathbf{P}_0$ denote the input model induced by $P_0$. The auxiliary input models $\mathbf{R}$ and $\mathbf{S}$ are given by

$$\mathbf{R}(o_t|o_{<t}) = \mathbf{P}_0(o_t|o_{<t}), \qquad \mathbf{S}(o_t|o_{<t}) = \mathbf{Pr}(o_t|o_{<t}).$$

Substituting these into (7.8) yields

$$\sum_{o_{\leq T}} \mathbf{P}_0(o_{\leq T})\mathbf{U}_*(o_{\leq T}) - \alpha \sum_{o_{\leq T}} \mathbf{P}_0(o_{\leq T}) \log \frac{\mathbf{P}_0(o_{\leq T})}{\mathbf{Pr}(o_{\leq T})}$$

Here, the solution is $\mathbf{P}(o_{\leq T}) = \mathbf{P}_0(o_{\leq T})$ for any utility function $\mathbf{U}_*$, and thus $\mathbf{P}(o_t|o_{<t})$ is the predictive distribution (Section 4.2.2)

$$\mathbf{P}(o_t|o_{<t}) = \sum_{\theta} P(\theta|o_{<t})P(o_t|o_{<t}),$$

as expected.

**Remark 21** The reference model $\mathbf{P}_0$ captures the current knowledge, which in this case happens to be a state of uncertainty induced by the Bayesian model $P_0$. This illustrates that Bayesian models do not add any further complications as they integrate seamlessly with the construction method. □

**Remark 22** This result looks trivial. However, notice that the variational problem for $o_t$ can be rewritten (dropping some constants that do not affect the variational problem) as

$$\mathbf{P}(o_t|o_{<t}) = \arg\min_{\mathbf{Pr}} \sum_{\theta} P(\theta) \sum_{o_{<t}} P(o_{<t}|\theta) \sum_{o_t} P(o_t|\theta, o_{<t}) \log \frac{P(o_t|\theta, o_{<t})}{\mathbf{Pr}(o_t|o_{<t})},$$

which is the standard way of formulating the problem. The question posed is: "what is the distribution $\mathbf{P}(o_t|o_{<t})$ that minimizes the average relative entropy to an unknown source that is chosen randomly?" This question is important in coding, because the $\mathbf{P}(o_t|o_{<t})$ chosen this way leads to the *optimal adaptive compressor*. From this point of view, the fact that the optimum is given by the predictive distribution (and hence the application of Bayes' rule) is not trivial at all. □

## 7.4 Historical Remarks & References

The formalization of bounded rationality presented in this chapter was entirely developed by the author and D. A. Braun (Ortega and Braun, 2010d). The conversion between probability

and utility in Section 7.2.1 has been introduced for the first time in Ortega and Braun (2010b). The proof of the variational principle is adapted from Keller (1998, Theorem 1.1.3).

Many of the presented concepts have pre-cursors in the literature. Previous theoretical studies have reported, for example, structural similarities between entropy and utility functions, see e.g. Candeal, De Miguel, Induráin, and Mehta (2001). A duality between optimal control and estimation has been reported by Todorov (2008), where an exponential transformation mediates between the cost-to-go function and a probability distribution that acts as a backwards filter. Free energy principles and the use of reference distributions for action selection have been proposed by Todorov (2009) and Kappen, Gomez, and Opper (2009). In these studies, transition probabilities of Markov systems were manipulated directly and the cost measured as a probabilistic deviation with respect to the passive dynamics of the system. The idea of bounded rationality through the consideration of information costs has been first proposed by Simon (1982). One of the first instantiations of bounded rationality was introduced in the context of risk in decision making (Borch, 1969). Accordingly, the decision maker prefers a gamble maximizing the mean utility penalized by the variance of the gamble, where the latter is multiplied by a risk factor. This is obviously conceptually in accordance with bounded SEU introduced here if we interpret variance as a *proxy for uncertainty*.

Besides the links to classical decision theory that the presented material has by design, there are obvious connections to information theory (Shannon, 1948), thermodynamics (see e.g. Callen 1985) and statistical inference (see e.g. maximum entropy principles, Jaynes and Bretthorst 2003). It remains to be seen whether these similarities are purely formal, or whether they rest on more fundamental principles.

# Chapter 8

# Causality

The maximum SEU principle assumes that an agent chooses its optimal policy before the interaction with its environment starts. As has been argued in Chapter 5 and further developed in Chapters 6 and 7, choosing the optimal policy requires running a computationally intractable optimization algorithm.

This assumption can be relaxed by letting the choice of the optimal policy happen *during interaction*, not before. This "dynamic choice" can be modeled by designing an agent that is initially uncertain about the optimal policy, but then learns it from the interactions with the environment. We can model this uncertainty over the optimal policy in the same way we modeled the uncertainty over the environment, introducing a Bayesian output model over a class of output models akin to the Bayesian input model. In this way, we can skip the costly optimization step and use a Bayesian mixture model to learn the optimal policy, driven exclusively by the interactions with the environment. Obviously, the cost we pay for doing so is that the resulting adaptive agent is suboptimal in the classical expected utility sense (although optimal in an information-theoretic sense that will be explained in Chapter 9).

However, this usage of Bayesian probability theory will violate a subtle but vital assumption underlying the reasoning model; namely, the predetermination of the outcome. If the agent is just a passive observer, then it does not matter whether the observations are generated on-the-fly or whether they were determined beforehand and then merely "uncovered" by the environment, because the formal treatment is the same in both cases. This is also true if the agent knows his policy. If the agent is certain about its policy and the policy is deterministic, then actions are just functions of the observations (Section 3.2.1). This in turn means that the I/O string can be modeled as if it were randomly chosen by the environment alone before the interaction starts, and then merely sequentially uncovered. However, if the agent is uncertain about its policy, then its actions are determined dynamically (say, by choosing them according to a stochastic rule), and hence not a function of the observation stream alone anymore. In this case the outcome is generated on-the-fly rather than predetermined: the agent chooses the actions and the environment chooses the observations.

Suppose there is an unknown cause influencing a result we are waiting for. As

soon as we observe the result, we learn something about the unknown cause. However, if instead we decide to interrupt the process by choosing the result ourselves (i.e. by acting), then our choice will not affect our knowledge about the unknown cause. *This is simply because we know that our current actions cannot change the past anymore.* Meanwhile, in both cases, we learn something about the future, i.e. about all the outcomes that will follow the result.

This distinction between belief updates following externally generated observations and internally generated actions is not modeled in Bayesian probability theory. Essentially, the theory lacks the formal tools to deal with indeterminate outcomes chosen by the reasoner himself. This requires introducing additional information to clearly identify the past and the future of choices, or more abstractly speaking, introducing a causal order of events. Inputs are incorporated into the knowledge by conditioning, while outputs have to be incorporated by first intervening the belief model and then conditioning.

The theory for enabling causal reasoning is known as *statistical causality*. Statistical causality is the study of the *functional dependencies* of events, i.e. the study of their cause-effect relationships. This stands in contrast to statistics proper, which, on an abstract level, can be said to study the *equivalence dependencies* of events, such as co-occurrence or correlation. Causal statements differ fundamentally from statistical statements. Examples that highlight the differences are many, such as "do smokers *get* lung cancer?" as opposed to "do smokers *have* lung cancer?" in a clinical study; "*assign* $y \leftarrow f(x)$" as opposed to "*compare* $y = f(x)$" in a programming language; "how does the increase of the price *change* the demand?" versus "how does the increase of the price *correlate with* the demand?" in econometrics; "does the legalization of abortion *decrease* the crime rate?" versus "does the legalization of abortion *relate to* the crime rate?" in the social sciences; and "$a \leftarrow F/m$" as opposed to "$F = m\,a$" in Newtonian physics.

Especially over the last decade, significant progress has been made towards the formal understanding of causation, mainly in the context of graphical models Pearl (2000); Spirtes and Scheines (2001); Dawid (2007), but also in the context of probability trees Shafer (1996). However, there is currently no set-theoretic formalization of causality (and interventions) that is naturally compatible with measure theory.

The aim of this chapter is to present a simple framework for causal reasoning that is compatible with the framework for uncertain reasoning introduced in Section 4.1.2. The axiomatization of causality presented in this chapter is entirely due to the author.

## 8.1 The Big Picture

The aim of this section is to give an overview of the concepts that are going to be formalized abstractly in the next section. Consider a three-stage experiment involving two identical glass urns: the left one containing one white and three black balls, and the right one having three white and one black ball. In stage one, the two urns are either swapped or not with equal probabilities. In stage two it is randomly decided

whether to exclude the left of the right urn from the experiment. If the urns have not been swapped in the first stage, then the odds are 3/4 and 1/4 for keeping the left and the right urn respectively. If the urns have been swapped, then the odds are reversed. In the last stage, a ball is drawn from the urn with equal probabilities and its color is revealed. We associate each stage with a binary random variable: namely Swap $\in$ {yes, no}, Pick $\in$ {left, right} and Color $\in$ {white, black} respectively. Figure 8.1 illustrates the setup using a probability tree (Shafer, 1996). In calculations, we will abbreviate variable names and their values with their first letters.

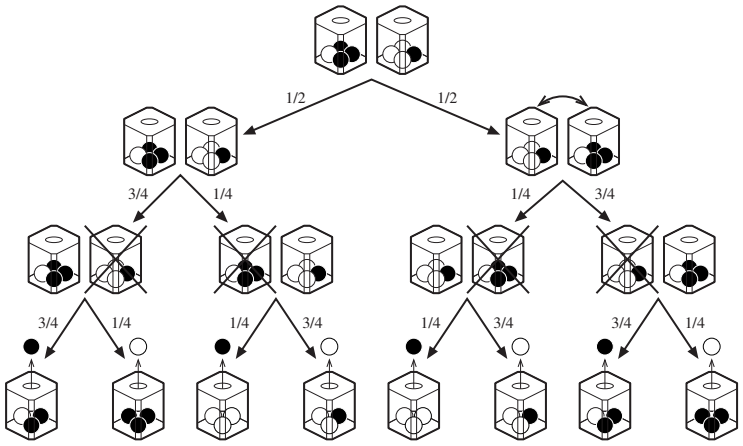



**Figure 8.1:** A Three-Stage Randomized Experiment.

The advantage of the representation as a probability tree is that it clearly highlights how the binary random variables Swap, Pick and Color are instantiated, including both the order and the probabilities. In this tree, the nodes corresponds to the (intermediate) steps of the realization of the experiment, where each node has a unique past (given path of nodes starting at the root node and ending at the current node) and many possible futures (given by any of the paths that follow from the node). The conditional probabilities of the probability tree fully determine a unique belief function. In particular, one can compute the probability of each realization (Table 8.1) by multiplying the conditional probabilities that make up the path of the realization.

Note that each level of the probability tree is associated with one and only one binary random variable. This is not very restrictive, as random variables taking on values in finite sets can always be broken down into sequences of binary random variables. Restricting ourselves to binary trees allows us representing each node as an intersection of "primitive" events. Informally, a primitive event tells us whether the realization took a left or a right branch at a given level of the tree.

The reason why we have chosen a probability tree representation of the experiment is because it clarifies the link between belief spaces introduced in Chapter 4 and Pearl's causal graphs (Pearl, 2000) which is currently one of the most-widely accepted formalisms for reasoning under **interventions**. The causal graph representation of the

**Table 8.1:** Probabilities of Realizations of the Experiment

| Swap | Pick | Color | Probability |
|---|---|---|---|
| no | left | black | 9/32 |
| no | left | white | 3/32 |
| no | right | black | 1/32 |
| no | right | white | 3/32 |
| yes | left | black | 1/32 |
| yes | left | white | 3/32 |
| yes | right | black | 9/32 |
| yes | right | white | 3/32 |

experiment is shown in Figure 8.2a. Intuitively, Pearl defines an intervention as the act of setting the value of a random variable, as opposed to passively observing it. In the graph, this operation amounts to removing the parent links to the node of the variable and then binding its value (Figure 8.2b). It turns out that the analogous operation in a probability tree corresponds to changing the conditional probabilities of a primitive event to either the left or the right branch (Figure 8.3). The next section introduces an abstract model of probability trees and interventions.

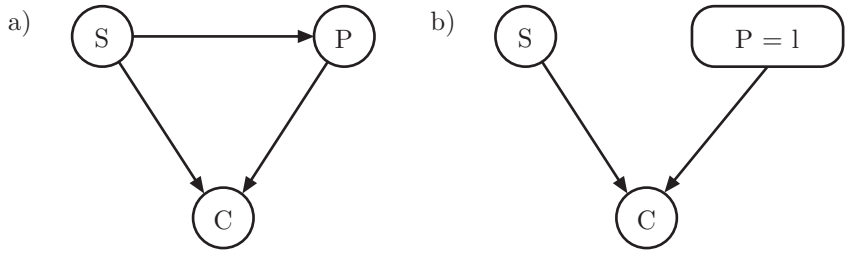



**Figure 8.2:** A Causal Graph. The graph shown in (a) is a representation of the experiment of Figure 8.1. Setting the value of the variable 'P' corresponds to an intervention, i.e. removing the parent link to the associated node and then recording the value of the random variable as shown in (b).

## 8.2 Causal Spaces

The aim of this section is to introduce causal spaces. Causal spaces contain enough information to characterize the causal structure of a random process.

Let $\Omega$ be a finite set of **outcomes**. An **atom set** $\mathcal{A}$ is a partition of $\Omega$, and an **atom** is a member $A \in \mathcal{A}$. Given a set $\mathcal{E}$ of subsets of $\Omega$, define the **algebra generated** by $\mathcal{E}$, written $\sigma(\mathcal{E})$, as the smallest algebra over $\Omega$ containing every member of $\mathcal{E}$.

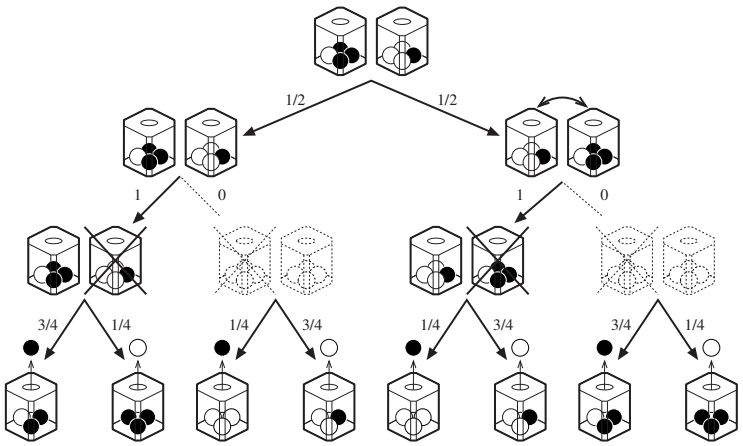



**Figure 8.3:** An Intervention in the Probability Tree. The figure illustrates choosing the left urn at the second stage of the experiment as in Figure 8.2b.

Furthermore, define the **atom set generated** by an algebra $\mathcal{F}$, written $\alpha(\mathcal{F})$, as the largest set of atoms containing members of $\mathcal{F}$. For any set $\mathcal{E}$ of subsets of $\Omega$, we also abbreviate $\alpha(\mathcal{E}) := \alpha(\sigma(\mathcal{E}))$.

**Remark 23** In the finite case, it is easily seen that both generated algebras and generated atom sets are unique. □

**Definition 25 (Primitive Events)** Let $E = (E_0, E_1, E_2, \ldots, E_N)$ be a finite sequence of subsets of $\Omega$ called **primitive events**, where $E_0 := \Omega$, and where for all $n \geq 1$,

$$E_n \notin \sigma\Big(\{E_0, E_1, \ldots, E_{n-1}\}\Big).$$

Furthermore, define $\mathcal{E}_n := \{E_n, E_n^c\}$ and $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_N$ as the sequence of atom sets

$$\mathcal{A}_n := \alpha\Big(\{E_0, E_1, \ldots, E_n\}\Big).$$

□

This setup is illustrated in Figure 8.4. The sequence of primitive events is an abstract characterization of a random process that occurs in discrete steps $n = 1, 2, \ldots, N$. Each step $n$ is associated with a primitive event $E_n$ representing a basic proposition whose truth value is resolved during this step (and not before!), i.e. step $n$ determines whether the outcome $\omega \in \Omega$ is either in $E_n$ or in $E_n^c$. The $n$-th atom set $\mathcal{A}_n$ contains one proposition for each possible path the random process can take. Therefore, after $n$ steps, the process will find itself in one (and only one) of the members in $\mathcal{A}_n$.

**Remark 24** The condition that $E_n$ cannot be in the algebra generated by the previous events $E_0, \ldots, E_{n-1}$ guarantees that $E_n$ adds a new proposition that cannot be expressed in terms of the previous propositions. □

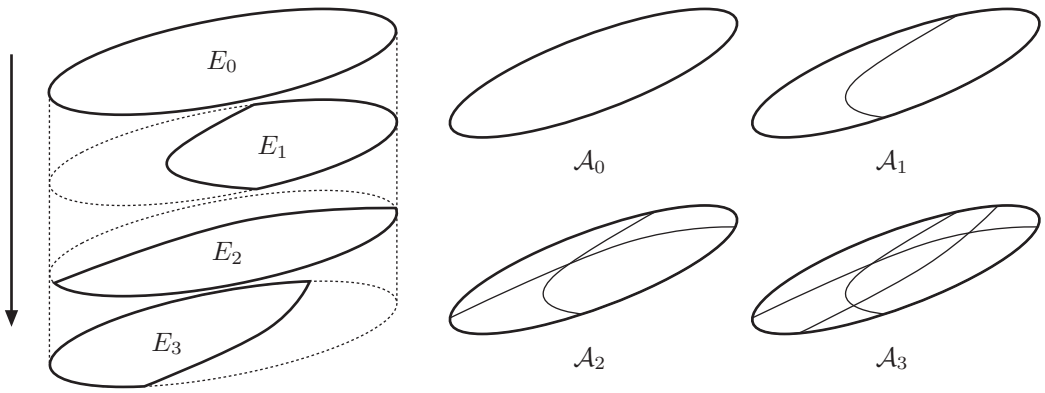



**Figure 8.4:** Primitive Events and their Atom Sets.

The sequence of primitive events $E = (E_1, \ldots, E_N)$ can equivalently be represented by any sequence $E' = (E'_1, \ldots, E'_N)$ where $E'_n \in \mathcal{E}_n$. Due to this, we will call any member of $\mathcal{E}_n$ primitive event. We introduce causal functions.

**Definition 26 (Causal Axioms)** Let $\Omega$ be a set of outcomes, and let $E = (E_1, \ldots, E_N)$ be a sequence of primitive events. A set function $\mathbf{C}_n$ is a **$n$-th causal function** iff

C1. $A \in \mathcal{E}_n, B \in \mathcal{A}_{n-1}, \quad \mathbf{C}_n(A|B) \in [0,1].$

C2. $A \in \mathcal{E}_n, B \in \mathcal{A}_{n-1}, \quad \mathbf{C}_n(A|B) = 1 \quad$ if $B \subset A$.

C3. $A \in \mathcal{E}_n, B \in \mathcal{A}_{n-1}, \quad \mathbf{C}_n(A|B) = 0 \quad$ if $A \cap B = \varnothing$.

C4. $A \in \mathcal{E}_n, B \in \mathcal{A}_{n-1}, \quad \mathbf{C}_n(A|B) + \mathbf{C}_n(A^c|B) = 1.$

Hence, $\mathbf{C}_n$ maps $\mathcal{E}_n \times \mathcal{A}_{n-1}$ into $[0,1]$. A **causal function** over $E$ is a function

$$\mathbf{C}(A|B) = \mathbf{C}_n(A|B), \qquad \text{if } A \in \mathcal{E}_n, B \in \mathcal{A}_{n-1},$$

where $\mathbf{C}_n$ is an $n$-th causal function. Hence, $\mathbf{C}$ maps $\bigcup_n (\mathcal{E}_n \times \mathcal{A}_{n-1})$ into $[0,1]$. □

The intuition behind this definition is as follows. The causal function specifies the knowledge the reasoner has about the evolution of a random process. It specifies the likelihood of a primitive event $A \in \mathcal{E}_n$ to happen after the random process is known to have taken a path $B \in \mathcal{A}_{n-1}$.

By comparing Axioms C1–C4 with Axioms B1–B5 (Section 4.1.2) of belief functions, we observe the following. First, in contrast to $\mathbf{B}$, only a subset of combinations $(A, B) \in \mathcal{F} \times \mathcal{F}$ is specified for $\mathbf{C}$, namely, the ones that chain a history of primitive events $B \in \mathcal{A}_{n-1} \subset \mathcal{F}$ together with the primitive event $A \in \mathcal{E}_n \subset \mathcal{F}$ that immediately follows. Second, Axioms C1–C4 play the same rôle as Axioms B1–B4, namely: (C1)

probabilities lie in the unit interval $[0, 1]$; (C2 & C3) probabilities are consistent with the truth function; and (C4) probabilities of complementary events add up to one. No axiom analogous to Axiom B5 is needed for $\mathbf{C}$.

Putting everything together, one gets a causal space. A causal space contains enough information to derive an associated belief space.

**Definition 27 (Causal Space)** A **causal space** is a tuple $(\Omega, E, \mathbf{C})$, where: $\Omega$ is a set of outcomes, $E$ is sequence of primitive events, and $\mathbf{C}$ is a causal function over $E$.□

**Definition 28 (Induced Belief Space)** Given a causal space $(\Omega, E, \mathbf{C})$, the **induced belief space** is the belief space $(\Omega, \mathcal{F}, \mathbf{B})$ where the algebra $\mathcal{F}$ and the belief function $\mathbf{B}$ are defined as

i. $\mathcal{F} = \sigma\Big(\{E_0, E_1, \ldots, E_N\}\Big)$;

ii. $\mathbf{B}(A|B) = \mathbf{C}(A|B)$, for all $(A, B) \in \bigcup_n (\mathcal{E}_n \times \mathcal{A}_{n-1})$. □

Thus, the induced belief space is constructed by generating the algebra $\mathcal{F}$ from the primitive events $E$, and by equating the belief function $\mathbf{B}$ to the causal function $\mathbf{C}$ over the subset of $\mathcal{F} \times \mathcal{F}$ where $\mathbf{C}$ is defined. The following theorem tells us that this subset is enough to completely determine the whole belief function.

**Theorem 7 (Induced Belief Space)** *The induced belief space exists and is unique.*□

Proof Let $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_N$ denote the sequence of algebras generated as

$$\mathcal{F}_n := \sigma(\{E_0, E_1, \ldots, E_n\}).$$

Let $r, s \in \mathbb{N}$, $r \leq s$, be the smallest numbers such that $B$ is $\mathcal{F}_r$-measurable and $A$ is $\mathcal{F}_s$-measurable. Let $\mathcal{B} \subset \mathcal{A}_r$ and $\mathcal{A} \subset \mathcal{A}_s$ be the partitions of $B$ and $A$ respectively.Then, $\mathbf{B}(A|B) = 0$ if $A \cap B$ by the belief axioms, and

$$\mathbf{B}(A|B) = \frac{\mathbf{B}(A|\Omega)}{\mathbf{B}(B|\Omega)}$$

otherwise. We have

$$\mathbf{B}(A|\Omega) = \sum_{a \in \mathcal{A}} \mathbf{B}(a|\Omega) \qquad \text{and} \qquad \mathbf{B}(B|\Omega) = \sum_{b \in \mathcal{B}} \mathbf{B}(b|\Omega).$$

For every $a \in \mathcal{A}$, let $a^1, a^2, \ldots, a^s$ the unique sequence $a^j \in \mathcal{E}_j$ such that

$$a = a^1 \cap a^2 \cap \cdots \cap a^s = \bigcap_{j=1}^{s} a^j.$$

Hence,

$$\mathbf{B}(a|\Omega) = \mathbf{B}\Big(\bigcap_{j=1}^{s} a^j \Big| \Omega\Big) = \prod_{j=1}^{s} \mathbf{B}\Big(a^j \Big| \Omega \cap \bigcap_{i=1}^{j-1} a^i\Big).$$

Similarly, one obtains for every $b \in \mathcal{B}$,

$$\mathbf{B}(b|\Omega) = \mathbf{B}\Big(\bigcap_{j=1}^{r} b^j \Big| \Omega\Big) = \prod_{j=1}^{r} \mathbf{B}\Big(b^j \Big| \Omega \cap \bigcap_{i=1}^{j-1} b^i\Big).$$

Thus, we have proven the following. First, $\mathcal{F}$ is unique because generated algebras are unique. Second, we have shown, for arbitrarily chosen events $A, B \in \mathcal{F}$, how to reexpress $\mathbf{B}(A|B)$ into an expression involving only terms of the form $\mathbf{C}(C|D)$. Hence, it cannot be that $\mathbf{B}, \mathbf{B}'$ are both consistent with $\mathbf{C}$ and there is $A, B \in \mathcal{F}$ such that $\mathbf{B}(A|B) \neq \mathbf{B}'(A|B)$. ■

### 8.2.1 Interventions

We now define the operation that specifies how the knowledge about the random process transforms when the reasoner himself intervenes it.

**Definition 29 (Intervention)** Given a causal space $(\Omega, E, \mathbf{C})$ and a primitive event $A \in \mathcal{E}_n$ for some $n \in \{1, \ldots, N\}$, the $A$-**intervention** is the causal space $(\Omega, E, \mathbf{C}')$ where for all $(B, C) \in \bigcup_n (\mathcal{E}_n \times \mathcal{A}_{n-1})$,

$$\mathbf{C}'(B|C) = \begin{cases} 1 & \text{if } A = B \text{ and } (B \cap C) \notin \{\varnothing, C\}, \\ 0 & \text{if } A = B^c \text{ and } (B \cap C) \notin \{\varnothing, C\}, \\ \mathbf{C}(B|C) & \text{else.} \end{cases}$$

□

This is an important definition. The reasoner ask himself the question: "How do my beliefs about the world change if I were to choose the truth value of a primitive event?" This is answered by *directly changing the causal function accordingly*. However, this change cannot contradict the logical constraints given by the underlying truth function.

**Remark 25** Note that $(B \cap C) \notin \{\varnothing, C\} \Leftrightarrow \mathbf{T}(B|C) = ?$. Hence, an intervention can only affect primitive propositions $B \in \mathcal{E}_n$ that have an unresolved truth value given the history $C \in \mathcal{A}_{n-1}$. Moreover, the intervention resolves the truth value of $B$. This makes sense intuitively. □

We will use the abbreviation $\hat{A}$ to denote $A$-interventions on a causal space. When the underlying causal space $(\Omega, E, \mathbf{C})$ inducing a belief space $(\Omega, \mathcal{F}, \mathbf{B})$ is clear from the context, then the expression $\mathbf{B}(B|\hat{A})$ denotes the belief $\mathbf{B}'(B|A)$ measured w.r.t. the belief space $(\Omega, \mathcal{F}, \mathbf{B}')$ induced by the $A$-intervention of $(\Omega, E, \mathbf{C})$. Furthermore, when $A \in \mathcal{F}$ is an event such that

$$A = \bigcap_{i=1}^{I} A_i,$$

where each $A_i$ is a primitive event, then the $A$-intervention is the causal space resulting as the succession of $A_i$-interventions.

## 8.3 Causality in Autonomous Systems

The previous section introduced causal spaces to model reasoning under uncertainty and control. The aim of this section is to apply this framework to model autonomous agents having uncertainty over their policies and their environments.

### 8.3.1 Bayesian I/O Model

We complete the description of the Bayesian model introduced in Section 4.2.1 by adding uncertainty over the policy.

**Definition 30 (Bayesian Output Model)** A **Bayesian output model** of an agent is a set of conditional probabilities

$$P(\theta), \quad P(a_t|\theta, \underline{ao}_{<t}) \quad \text{for all} \quad (\theta, \underline{ao}_{\leq t}) \in \Theta \times \mathcal{Z}^\diamond.$$

inducing a unique probability measure $P$ over $\Theta \times \mathcal{A}^T$ conditioned on $\mathcal{O}^T$. □

**Definition 31 (Bayesian I/O Model)** A **Bayesian I/O model** of an agent is a Bayesian input model paired with a Bayesian output model, i.e. a set of conditional probabilities

$$P(\theta), \quad P(a_t|\theta, \underline{ao}_{<t}) \quad \text{and} \quad P(o_t|\theta, \underline{ao}_{<t} a_t) \quad \text{for all} \quad (\theta, \underline{ao}_{\leq t}) \in \Theta \times \mathcal{Z}^\diamond.$$

inducing a unique probability measure $P$ over $\Theta \times \mathcal{Z}^T$. □

The intuition is analogous to the Bayesian input model. The set $\Theta$ is a collection of **unknown parameters** that completely specify the I/O model, i.e. knowing $\theta$ would lead to a unique I/O model given by

$$\begin{aligned} \mathbf{P}_\theta(a_t|\underline{ao}_{<t}) &= P(a_t|\theta, \underline{ao}_{<t}) \\ \mathbf{P}_\theta(o_t|\underline{ao}_{<t} a_t) &= P(o_t|\theta, \underline{ao}_{<t} a_t) \quad \text{for all} \quad \underline{ao}_{\leq t} \in \mathcal{Z}^\diamond. \end{aligned}$$

### 8.3.2 Causal Structure

As has been pointed out in the introduction to this chapter, when using a model of beliefs, one has to distinguish between inputs (provided externally by the environment) and outputs (generated internally). Inputs are incorporated into the knowledge by conditioning, while outputs have to be incorporated by first applying the corresponding intervention to the belief model and then conditioning.

The Bayesian I/O model (and all the models of autonomous systems we have introduced so far) is specified by conditional probabilities over the next symbol given its past. It is therefore actually a specification of the causal structure! More precisely, consider *any* causal space $(\Omega, E, \mathbf{C})$ fulfilling the following requirements. First, the primitive events are given by

$$E = \Big(E_0, \underbrace{E_1, \ldots, E_{N_\theta}}_{\theta}, \underbrace{E_{N_\theta+1}, \ldots, E_{N_{a_1}}}_{a_1}, \underbrace{E_{N_{a_1}+1}, \ldots, E_{N_{o_1}}}_{o_1}, \ldots, \underbrace{E_{N_{a_T}+1}, \ldots, E_{N_{o_T}}}_{o_T}\Big),$$

where each of the highlighted blocks of primitive events generates a partition of $\Omega$ containing enough members for specifying a random variable (i.e. either $\theta$, $a_t$ or $o_t$ for $t = 1, \ldots, T$). Note that $N_\theta \leq N_{a_1} \leq N_{o_1} \leq \cdots \leq N_{o_T} = N$. This is illustrated in Figure 8.5. Choosing an action $a_t$ corresponds to an $a_t$-intervention (that breaks down into a sequence of $E_n$-interventions). Second, the causal function $\mathbf{C}$ is such that the induced belief function $\mathbf{B}$ coincides with the Bayesian model, i.e. for all $(\theta, \underline{ao}_{\leq t}) \in \Theta \times \mathcal{Z}^\diamond$,

$$\begin{aligned} \mathbf{B}(\theta) &= P(\theta), \\ \mathbf{B}(a_t|\theta, \underline{ao}_{<t}) &= P(a_t|\theta, \underline{ao}_{<t}) \\ \text{and} \quad \mathbf{B}(a_t|\theta, \underline{ao}_{<t} a_t) &= P(o_t|\theta, \underline{ao}_{<t} a_t). \end{aligned}$$

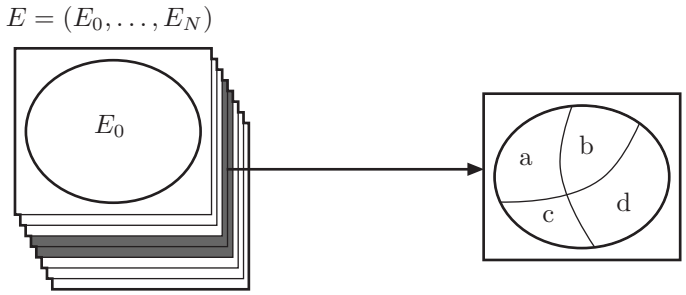



**Figure 8.5:** Causal Space of an Autonomous System. The sequence of primitive events is represented as a stack of sets on the left hand side. The figure shows how a random variable (in this case having the alphabet {a,b,c,d}) is constructed from two successive primitive events.

### 8.3.3 Belief Updates

The correspondence between causal spaces and Bayesian I/O models has the following operational implications. When an autonomous system interacts with the environment, then its belief state is updated. This update depends on whether the symbol is an input or an output. If it is an input, then the update is a condition. If it is an output, then the update is an intervention followed by a condition. This difference is illustrated in Figure 8.6.

Assume there is a sequence of random variables $X_1, X_2, \ldots, X_T$ taking on values in $\mathcal{X}$ with conditional probability measures $\mathbf{B}(X_t|X_{<t})$, $t = 1, \ldots, T$.

An observation is a **measurement**. As such, it provides information about the whole realization of the stochastic process. That is, learning the value of $X_t$ provides information about all $\{X_s : 1 \leq s \leq T\}$ through the dependencies established by the causal model. The update $X_t = x_t$ following an observation changes all conditional probabilities as

$$\mathbf{B}(A|B) \quad \xrightarrow{X_t = x_t} \quad \mathbf{B}(A|B, X_t = x_t),$$

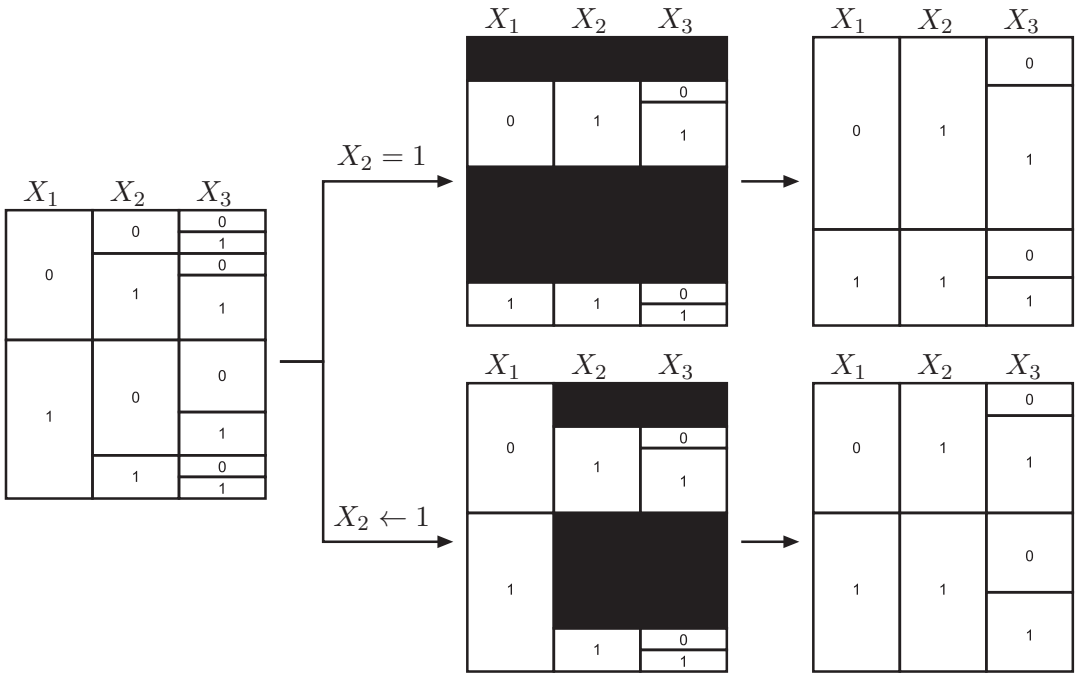



**Figure 8.6:** Updates following an Observation versus an Action. The figure shows three causally ordered random variables $X_1$, $X_2$ and $X_3$ (taking on binary values) and their probabilities (through the height of their boxes). Two updates are compared: the update $X_2 = 1$ following an observation and the update $X_2 \leftarrow 1$ following an action. These updates eliminate the incompatible probability mass (as shown in the first column after the update) and then normalize the remaining probability mass (second column after the update). The observation affects the probability mass of the whole history, eliminating the incompatible realizations and then normalizing globally. The action affects only the probability mass of the present and the future and then normalizes within each past.

where $A$ and $B$ are arbitrary events.

An action is a **decision**. As such, it only provides information about the future of the realization of the stochastic process, but not about its past. That is, learning the value of $X_t$ provides information about $\{X_s : t \leq s \leq T\}$ *only*. The update $X_t \leftarrow x_t$ following an action changes all conditional probabilities as

$$\mathbf{B}(A|B) \quad \xrightarrow{X_t \leftarrow x_t} \quad \mathbf{B}(A|B, X_t \leftarrow x_t) = \mathbf{B}'(A|B, X_t = x_t),$$

where $A$ and $B$ are arbitrary events and where $\mathbf{B}'$ is the probability measure uniquely defined by the equations

$$\begin{array}{lll} \text{i.} & \mathbf{B}'(X_{<t}) = \mathbf{B}(X_{<t}), & \text{(past)} \\ \text{ii.} & \mathbf{B}'(X_t|X_{<t}) = \delta_{x_t}(X_t), & \text{(present)} \\ \text{iii.} & \mathbf{B}'(X_{t+1:T}|X_{\leq t}) = \mathbf{B}(X_{t+1:T}|X_{\leq t}). & \text{(future)} \end{array} \tag{8.1}$$

When the random variable $X_t$ is clear from the context, we use the abbreviation

$$\mathbf{B}(A|B, \hat{x}_t) := \mathbf{B}(A|B, X_t \leftarrow x_t).$$

Hence, when an autonomous system (having a Bayesian I/O model $P$) experiences the I/O string $\underline{ao}_{\leq t} \in \mathcal{Z}^\diamond$, then its probabilities are given by

$$P(A|\underline{\hat{a}o}_{\leq t}),$$

where $A$ is an arbitrary event. It is easily shown that

$$P(a_t|\theta, \underline{\hat{a}o}_{<t}) = P(a_t|\theta, \underline{ao}_{<t}) \quad \text{and} \quad P(o_t|\theta, \underline{\hat{a}o}_{<t}\hat{a}_t) = P(o_t|\theta, \underline{ao}_{<t}a_t).$$

**Remark 26** In general, one has that $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t) \neq P(o_t|\underline{ao}_{<t}a_t)$. Note however, that when the policies of the Bayesian model are all the same, that is, $P(a_t|\theta, \underline{\hat{a}o}_{<t}) = P(a_t|\theta', \underline{\hat{a}o}_{<t})$ for all $\theta, \theta' \in \Theta$ (or, if one prefers, there is *no uncertainty* over the policy), then $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t) = P(o_t|\underline{ao}_{<t}a_t)$. *This clarifies why causality does not play a role when systems are constructed using the maximum SEU principle.* □

### 8.3.4 Induced I/O Model

We have seen that the Bayesian I/O model allows modeling both the uncertainty over the policy and the predictor of an autonomous system.

**Definition 32 (Induced I/O Model)** The I/O model **P induced** by a Bayesian I/O model $P$ is defined as

$$\mathbf{P}(a_t|\underline{ao}_{<t}) := P(a_t|\underline{\hat{a}o}_{<t}) \quad \text{and} \quad \mathbf{P}(o_t|\underline{ao}_{<t}a_t) := P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$$

for all $\underline{ao}_{<t} \in \mathcal{Z}^\diamond$. □

Before we give an intuitive interpretation of the induced model, we want to find out what the quantities $P(a_t|\underline{\hat{a}o}_{<t})$ and $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$ actually are.

**Theorem 8 (Induced I/O Model)** *The quantities* $P(a_t|\underline{\hat{a}o}_{<t})$ *and* $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$ *are given by*

$$P(a_t|\underline{\hat{a}o}_{<t}) = \sum_{\theta} P(a_t|\theta, \underline{ao}_{<t})P(\theta|\underline{\hat{a}o}_{<t}) \tag{8.2}$$

$$\textit{and} \quad P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t) = \sum_{\theta} P(o_t|\theta, \underline{ao}_{<t}a_t)P(\theta|\underline{\hat{a}o}_{<t}), \tag{8.3}$$

*where* $P(\theta|\underline{\hat{a}o}_{\leq t})$ *is given by*

$$P(\theta|\underline{\hat{a}o}_{\leq t}) = \frac{P(o_t|\theta, \underline{ao}_{<t}a_t)P(\theta|\underline{\hat{a}o}_{<t})}{\sum_{\theta'} P(o_t|\theta, \underline{ao}_{<t}a_t)P(\theta'|\underline{\hat{a}o}_{<t})} \qquad \textit{(recursion)} \tag{8.4}$$

$$= \frac{P(\theta)\prod_{\tau=1}^{t} P(o_\tau|\theta, \underline{ao}_{<\tau}a_\tau)}{\sum_{\theta'} P(\theta')\prod_{\tau=1}^{t} P(o_\tau|\theta', \underline{ao}_{<\tau}a_\tau)}. \qquad \textit{(function)} \tag{8.5}$$

□

Proof For $P(a_t|\underline{\hat{a}o}_{<t})$, introduce the variable $\theta$ via marginalization and then applying the chain rule:

$$P(a_t|\underline{\hat{a}o}_{<t}) = \sum_{\theta} P(a_t|\theta, \underline{\hat{a}o}_{<t})P(\theta|\underline{\hat{a}o}_{<t}).$$

First, note that $P(a_t|\theta, \underline{\hat{a}o}_{<t}) = P(a_t|\theta, \underline{ao}_{<t})$. Then, the term $P(\theta|\underline{\hat{a}o}_{<t})$ can be further expanded as

$$\begin{aligned} P(\theta|\underline{\hat{a}o}_{<t}) &= \frac{P(\underline{\hat{a}o}_{<t}|\theta)P(\theta)}{\sum_{\theta'} P(\underline{\hat{a}o}_{<t}|\theta')P(\theta')} \\ &= \frac{P(\theta)\prod_{\tau=1}^{t-1} P(\hat{a}_\tau|\theta, \underline{\hat{a}o}_{<\tau})P(o_\tau|\theta, \underline{\hat{a}o}_{<\tau}\hat{a}_\tau)}{\sum_{\theta'} P(\theta')\prod_{\tau=1}^{t-1} P(\hat{a}_\tau|\theta', \underline{\hat{a}o}_{<\tau})P(o_\tau|\theta', \underline{\hat{a}o}_{<\tau}\hat{a}_\tau)} \\ &= \frac{P(\theta)\prod_{\tau=1}^{t-1} P(o_\tau|\theta, \underline{ao}_{<\tau}a_\tau)}{\sum_{\theta'} P(\theta')\prod_{\tau=1}^{t-1} P(o_\tau|\theta', \underline{ao}_{<\tau}a_\tau)}. \end{aligned}$$

The first equality is obtained by applying Bayes' rule and the second by using the chain rule for probabilities. The second equality follows from using the causal factorization of the joint probability distribution. To get the last equality, one applies the interventions to the causal factorization. Thus, $P(\hat{a}_\tau|\theta, \underline{\hat{a}o}_{<\tau}) = 1$ and $P(o_\tau|\theta, \underline{\hat{a}o}_{<\tau}\hat{a}_\tau) = P(o_\tau|\theta, \underline{ao}_{<\tau}a_\tau)$. The recursive characterization of $P(\theta|\underline{ao}_{\leq t})$ is obtained trivially following similar arguments and by applying Bayes' rule only over $\underline{ao}_t$. Finally, the equations characterizing $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$ are derived analogously. ■

In the Bayesian I/O model, one assumes that the autonomous system has uncertainty over the unknown parameter $\theta \in \Theta$ that captures all the relevant information to completely define an I/O model. More precisely, it is assumed that $\theta$ is going to be drawn with probability $P(\theta)$, thus defining: $P(o_t|\theta, \underline{ao}_{<t}a_t)$, the true input model characterizing the true environment; and $P(a_t|\theta, \underline{ao}_{<t})$, the true policy that *should be applied*.

**Remark 27** Recalling that $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$ implements the predictive distribution and comparing it to $P(a_t|\underline{\hat{a}o}_{<t})$, it is reasonable to ask whether the latter is a "predictive output distribution" having analogous properties, i.e. a distribution over the outputs that converges to the "true optimal policy". Experimental evidence suggests that this might be the case. This will be further investigated in Chapter 9. □

## 8.4 Historical Remarks & References

The axiomatization of causality presented in this chapter is entirely due to the author. This formalization is based on previous work by other authors. The importance of distinguishing between input and output, more commonly known as the difference between *seeing and doing*, and their impact on inference, has been mainly developed by Pearl (2000) in the context of graphical models. Especially, Pearl introduced the formalization of interventions in so-called causal graphs and extended them later to the context of structural equations. Furthermore, first ideas of how to formalize the causal structure of a random process in set-theoretic terms were analyzed by Shafer (1996), although his formalization does not clarify how to extend it to interventions. Of particular relevance for the context of autonomous systems is *causal decision theory* (Stalnacker, 1968; Nozick, 1969; Lewis, 1973). This theory maintains that the expected utility of actions should be evaluated with respect to their potential causal consequences. Essentially, this proposes that there is a difference between actions and observations.

The study of causality has recently enjoyed considerable attention from the researchers in the fields of statistics and machine learning. Especially over the last decade, significant progress has been made towards the formal understanding of causation. For a more in-depth exposition of causality and its applications, the reader is referred to the specialized literature. For instance: Suppes (1970); Cartwright (1989); Eells (1991); Mellor (1995); Shafer (1996); Pearl (2000); Spirtes and Scheines (2001); and Dawid (2007).

# Chapter 9

# Control as Estimation

Bounded SEU introduced in Chapter 7 allows conceptualizing autonomous agents under bounded information-theoretic resources. These can be obtained by means of two possible transformations (or a combination of them): the control transformation and the estimation transformation. The control transformation solves the problem of constructing an autonomous agent optimizing the expected utility under resource constraints. The estimation transformation solves the problem of constructing an autonomous agent approximating a "known" reference autonomous agent under resource constraints.

Chapter 8 introduced a Bayesian model of autonomous systems that allows representing the uncertainty over the optimal policy and over the environment. Furthermore, it explained how to carry out the belief updates following inputs and following outputs, emphasizing the different causal constraints they are subject to.

*The Bayesian model and the estimation transformation can be used jointly to formulate an autonomous agent that "discovers" the optimal policy during the interactions with the environment.* That is, we can formulate an autonomous agent whose policy at the behavioral level can be interpreted as having uncertainty over a set of policies at the belief level. Then, while the agent is interacting with the environment, it obtains evidence that progressively eliminates this uncertainty. This is in stark contrast to autonomous agents that are constructed following the maximum SEU principle, where only one policy (namely, the optimal) is chosen right from the beginning, even before the interaction with the environment has even started.

One of the main points that has been put forward in this thesis is that obtaining certainty is expensive, be it through computation or interaction with the environment. Furthermore, it has been argued at a number of places in this thesis that computing information oneself is more expensive than being told the same information. Recall that in Chapter 6 and 7 we have argued that the resource costs spent in a transformation can be measured by the relative entropy between the initial and the final distribution. Intuitively speaking, the advantage of the method presented in this chapter lies in the fact that we do not have to pre-compute the optimal policy. Instead, we can already design an agent using just the distribution over policies that best describes our prior knowledge. Hence, instead of having to transform the distribution over the optimal

policy before the interaction starts, we can slowly transform this distribution guided by the observations alone.

## 9.1 Interlude: Dynamic versus Static

During the course of this thesis, we have spoken on various occasions about dynamic policies as opposed to pre-computed or static policies. Similarly, we have spoken about stochastic and deterministic policies. What does this mean, what is the difference and what are the implications?

Consider the well-known rock-paper-scissors game. In this game, non-random behavior can be exploited, and thus the safest way to protect oneself against an opponent consists in playing with uniform probability[1]. Suppose we have to play a sequence of rock-paper-scissors games. The uniform strategy can be implemented statically, e.g. by sampling each move beforehand and revealing them turn by turn; or, it can be implemented dynamically, e.g. by deciding each move just before it has to be played[2]. This flexibility does not exists if we wanted to use a deterministic strategy, because by choosing a deterministic policy we are determining all the moves beforehand. In general, having uncertainty allows us determining the value of random variables dynamically.

### 9.1.1 Risk versus Ambiguity

A special case occurs when there is uncertainty over the very beliefs of the decision maker, as is the case modeled by the Bayesian I/O model introduced in 3.2.2. The author proposes that this kind of uncertainty relates to an old debate dating from the early conceptualizations of decision theory: **risk** versus **ambiguity** (Knight, 1921).

Risk corresponds to the odds that the decision maker has adequate knowledge of, whereas ambiguity to the odds that are unknown to him. From the two, risk is well-understood; indeed, classical decision theory was regarded by its very founders as a "normative foundation of optimal decision making under risk" (von Neumann and Morgenstern, 1944; Savage, 1954). The success of these frameworks had put under question the operational relevance of the concept of ambiguity, until Ellsberg presented the paradox named after him in his seminal paper (Ellsberg, 1961). Several mathematical formalizations for ambiguity have been proposed in the literature, but none of them has found widespread acceptance (Camerer and Weber, 1992). However, on an abstract level the vast majority of approaches considers ambiguity to be some kind of uncertainty about risk or uncertainty of belief or lack of information. Here we propose a simple formalization that is an original contribution of this thesis.

To illustrate the formalization, let us review an example. Imagine a rational decision maker has to place a bet over the outcome of a biased coin toss which is either Head

[1] In game-theoretic parlance, it is said that the rock-paper-scissors game has a mixed strategy **Nash equilibrium** (Osborne and Rubinstein, 1999).

[2] In computer science, this technique of delaying the evaluation of a quantity up until the point where it is needed is known as *lazy evaluation* (Pratt and Zelkowitz, 2000).

or Tail. The payoff is \$1 for a correct bet or \$0 for a wrong bet. Given the rational decision maker's belief, we want to predict the bet he will place under five different cases, illustrated in Figure 9.1:

I. He believes that the odds are $\frac{1}{4}$ for Head and $\frac{3}{4}$ for Tail.

II. He believes that the odds are $\frac{3}{4}$ for Head and $\frac{1}{4}$ for Tail.

III. He believes that the odds $\frac{5}{8}$ for Head and $\frac{3}{8}$ for Tail.

IV. He believes that either I or II occurs with probability $\frac{1}{4}$ or $\frac{3}{4}$ respectively.

V. He *finds himself believing* in either I or II with probability $\frac{1}{4}$ or $\frac{3}{4}$ respectively.

Cases I–IV are easily examined under the framework of maximum expected utility. The rational decision maker places the bet that maximizes his expected payoff. For cases I and II, the optimal bets are Tail and Head respectively, because their expected payoff is \$0.75, as opposed to \$0.25 offered by the alternative bet (Figures 9.1a & b). Likewise, in III the decision maker bets Head. In case IV, the decision maker can reexpress the two-stage coin toss, i.e. first selecting between situation I or II and then tossing the coin (Figure 9.1c), as an equivalent single-stage coin toss with a re-weighted bias obtained by multiplying the probabilities of the first stage with the probabilities of the second stage (Figure 9.1e). This reduction reveals that case IV is equivalent to case III, with Head being to optimal bet.

However, by construction, case V requires a different analysis. Here, the decision maker's belief can take on one out of two possible forms, in which the optimal bets are Tail and Head as discussed previously. *The crucial difference lies in the fact that the probabilities of the belief instantiations are beyond the scope of the decision maker's analysis.* Therefore, for him it is optimal to bet Tail when reaching case I and to bet Head when reaching case II. This innocuous fact is by no means trivial, because the subjective expected utility of the decision maker is $\frac{1}{4} \cdot \frac{3}{4} + \frac{3}{4} \cdot \frac{3}{4} = \frac{3}{4} > \frac{5}{8}$, that is, *strictly higher* than the subjective expected utility of the classical analysis—and in practice, subjective beliefs are all what a decision maker has[3]! We call these probabilities that are exogenous to the decision maker's beliefs *ambiguities*.

The distinction has an operational meaning. Notice that in case V, the belief of the decision maker is itself a random variable, implying that the optimal policy is undefined until the random variable is resolved. Hence, the computation of the optimal policy can be delayed, i.e. the optimal policy can be determined dynamically. This is unlike case IV, where the policy is pre-computed/static. The corresponding Bayesian I/O model is as follows. Let $\theta \in \{\text{I, II}\}$ be the parameter determining whether the decision maker

[3]Intuitively, it seems safer to delay one's decision until the evidence is conclusive.

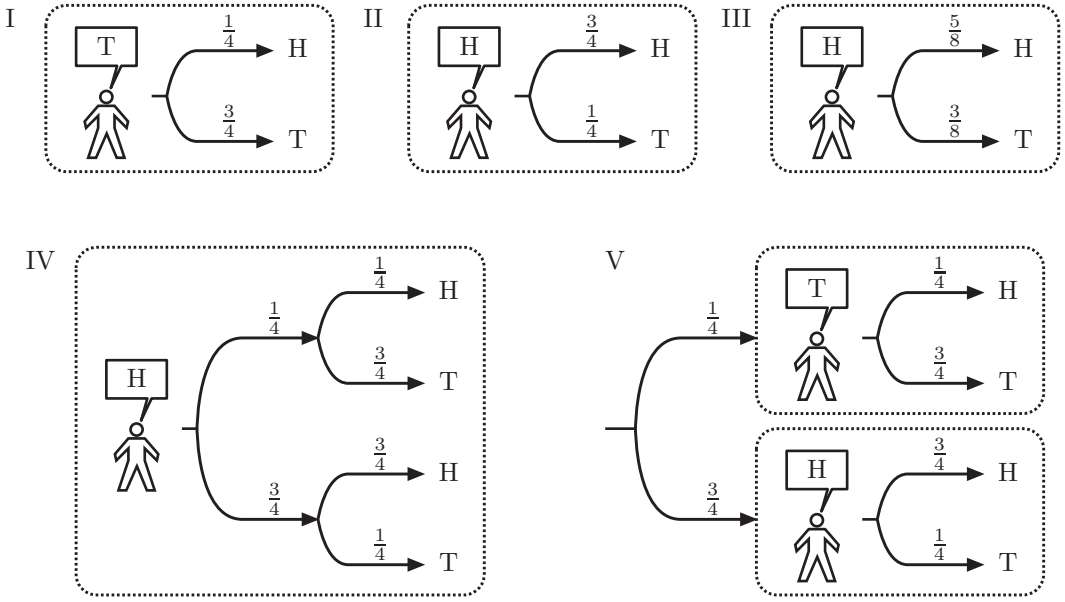



**Figure 9.1:** Risk versus Ambiguity. In the figure, five different decision making scenarios are shown. A biased coin is tossed. The goal is to predict the outcome and the payoffs are \$1 and \$0 for a right and wrong guess respectively. A rational decision maker places bets (here shown inside speech bubbles) such that his subjective expected utility is maximized. These subjective beliefs are delimited within dotted boxes. The cases in panels I–IV differ from case V in that the former can be fully understood in terms of classical decision theory, whereas the latter cannot.

is in case I or II. Then, the full Bayesian I/O model is given by:

$$P(\theta = \mathrm{I}) = \frac{1}{4} \qquad P(\theta = \mathrm{II}) = \frac{3}{4}$$
$$P(a|\theta = \mathrm{I}) = \begin{cases} 0 & \text{if } a = \mathrm{H} \\ 1 & \text{if } a = \mathrm{T} \end{cases} \qquad P(a|\theta = \mathrm{II}) = \begin{cases} 1 & \text{if } a = \mathrm{H} \\ 0 & \text{if } a = \mathrm{T} \end{cases}$$
$$P(o|\theta = \mathrm{I}, a) = \begin{cases} \frac{1}{4} & \text{if } o = \mathrm{H} \\ \frac{3}{4} & \text{if } o = \mathrm{T} \end{cases} \qquad P(o|\theta = \mathrm{II}, a) = \begin{cases} \frac{3}{4} & \text{if } o = \mathrm{H} \\ \frac{1}{4} & \text{if } o = \mathrm{T}. \end{cases}$$

Here, the prior probabilities $P(\theta)$ are ambiguities while the $P(a|\theta)$ and $P(o|\theta, a)$ are risk probabilities, because fixing $\theta$ determines the decision maker's estimation about the outcome and his policy.

## 9.2 Adaptive Estimative Control

If the environment $\mathbf{Q}$ is *unknown*, then the task of designing an appropriate agent $\mathbf{P}$ constitutes an *adaptive control problem*. We have already seen how to solve such a problem using the maximum SEU principle in Section 4.2, where it has been formulated as an *adaptive optimal control problem*.

We formulate a different problem setup that will be termed **adaptive estimative control**. Specifically, this setup deals with the case when the designer has a class of output models $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ parameterized by a finite set $\Theta$, designed to fit to a class of environments $\{\mathbf{Q}_\theta\}_{\theta\in\Theta}$; in other words, *when the designer wants to use policy* $\mathbf{P}_\theta(a_t|\underline{ao}_{<t})$ *for environment* $\mathbf{Q}_\theta(o_t|\underline{ao}_{<t}a_t)$.

Formally, it is assumed that we have a reference Bayesian I/O model (Section 8.3.1) characterized by conditional probabilities

$$P_0(a_t|\theta, \underline{ao}_{<t}) = \mathbf{P}_\theta(a_t|\underline{ao}_{<t}) \quad \text{and} \quad P_0(o_t|\theta, \underline{ao}_{<t}a_t) = \mathbf{P}_\theta(o_t|\theta, \underline{ao}_{<t}a_t)$$

representing $\theta$-th I/O model, and by probabilities $P(\theta)$ over the unknown parameter $\theta \in \Theta$. We will call the set of I/O models indexed by $\Theta$ the **set of operation modes** and a particular I/O model an **operation mode**. The objective is to find an I/O model $\mathbf{P}$ obtained by maximizing the bounded SEU using an estimation transformation to solve for all its input and outputs.

## 9.3 Bayesian Control Rule

One can derive a solution to the adaptive estimative control problem. Given the reference Bayesian I/O model $P_0$, let $\mathbf{P}_0$ denote the induced I/O model, i.e. the I/O model defined by (Section 8.3.4)

$$\mathbf{P}_0(a_t|\underline{ao}_{<t}) = P_0(a_t|\hat{a}\underline{o}_{<t}) \quad \text{and} \quad \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) = P_0(o_t|\hat{a}\underline{o}_{<t}\hat{a}_t).$$

Then, define the auxiliary I/O models $\mathbf{R}$ and $\mathbf{S}$ as

$$\begin{aligned} \mathbf{R}(a_t|\underline{ao}_{<t}) &= \mathbf{P}_0(a_t|\underline{ao}_{<t}), & \mathbf{R}(o_t|\underline{ao}_{<t}a_t) &= \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t), \\ \mathbf{S}(a_t|\underline{ao}_{<t}) &= \mathbf{Pr}(a_t|\underline{ao}_{<t}), & \mathbf{S}(o_t|\underline{ao}_{<t}a_t) &= \mathbf{Pr}(o_t|\underline{ao}_{<t}a_t). \end{aligned}$$

Substituting these into (7.8) yields

$$\sum_{\underline{ao}_{\leq T}} \mathbf{P}_0(\underline{ao}_{\leq T})\mathbf{U}_*(\underline{ao}_{\leq T}) - \alpha \sum_{\underline{ao}_{\leq T}} \mathbf{P}_0(\underline{ao}_{\leq T}) \log \frac{\mathbf{P}_0(\underline{ao}_{\leq T})}{\mathbf{Pr}(\underline{ao}_{\leq T})}$$

Here, the solution is $\mathbf{P}(\underline{ao}_{\leq T}) = \mathbf{P}_0(\underline{ao}_{\leq T})$ for any utility function $\mathbf{U}_*$. The policy and the predictor are thus given by $\mathbf{P}(a_t|\underline{ao}_{<t})$ and $\mathbf{P}(o_t|\underline{ao}_{<t}a_t)$ respectively. According to Theorem 8, the input model is given by the predictive distribution

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) = \sum_\theta P_0(\theta|\underline{\hat{a}o}_{<t}) P_0(o_t|\theta, \underline{ao}_{<t}a_t).$$

The associated output model is given by a quantity which we now promote to a definition.

**Proposition 2 (Bayesian Control Rule)** *The probability of $a_t$ conditioned on the past $\underline{ao}_{<t}$ is given by*

$$\mathbf{P}(a_t|\underline{ao}_{<t}) = \sum_\theta P_0(\theta|\underline{\hat{a}o}_{<t}) P_0(a_t|\theta, \underline{ao}_{<t}) \tag{9.1}$$

*where the mixture weights $P(\theta|\underline{ao}_{\leq t})$ are given by the recursion*

$$P_0(\theta|\underline{\hat{a}o}_{\leq t}) = \frac{P_0(o_t|\theta, \underline{ao}_{<t}a_t) P_0(\theta|\underline{\hat{a}o}_{<t})}{\sum_{\theta'} P_0(o_t|\theta, \underline{ao}_{<t}a_t) P_0(\theta'|\underline{\hat{a}o}_{<t})}.$$

*We call Equation (9.1) the **Bayesian control rule**.* □

**Remark 28** Note that the Bayesian control rule is a *stochastic control law*, i.e. actions $a_t$ are *sampled* from this distribution. □

Essentially, the previous result tells us that the Bayesian I/O model is the solution to the adaptive estimative control problem. This result leads to a family of algorithms that is very easy to implement in practice. The essential idea of the Bayesian control rule is to regard $P_0(a_t|\theta, \underline{ao}_{<t})$ as the policy to apply when the agent faces the environment described by $P_0(o_t|\theta, \underline{ao}_{<t}o_t)$, and then to perform a Bayesian (posterior) mixture over all $\theta \in \Theta$. Then, at each time step, one samples a parameter $\theta$ from the posterior distribution $P(\theta|\underline{\hat{a}o}_{\leq t})$ and then executes the policy indexed by $\theta$.

## 9.4 Convergence of the Bayesian Control Rule

In this section, let $\mathbf{P}$ denote a I/O model induced by a Bayesian I/O model $P$. We have seen that the predictive distribution converges to the true input stream (Section 4.2.4). Can something similar be said about the Bayesian control rule? That is, is it true that the Bayesian control rule converges to the "correct" output model, i.e.

$$P(a_t|\underline{\hat{a}o}_{<t}) \xrightarrow{t\to\infty} P(a_t|\theta_*, \underline{ao}_{<t}),$$

where $\theta_* \in \Theta$ is such that $P(o_t|\theta_*, \underline{ao}_{<t}a_t) = \mathbf{Q}(o_t|\theta_*, \underline{ao}_{<t}a_t)$? In this chapter we sketch a proof for a very restricted setting and we present experimental evidence that this might be the case in a more general setting.

### 9.4.1 Policy Diagrams

In the following we use "policy diagrams" as a useful informal tool to analyze the effect of policies on environments. One can imagine an environment as a collection of states connected by transitions labeled by I/O symbols (Figure 9.2). The zoom highlights a state $s$ where taking action $a \in \mathcal{A}$ and collecting observation $o \in \mathcal{O}$ leads to state $s'$. Sets of states and transitions are represented as enclosed areas similar to a Venn diagram. Choosing a particular policy in an environment amounts to partially controlling the transitions taken in the state space, thereby choosing a probability distribution over state transitions (e.g. a Markov chain given by the environmental dynamics). If the probability mass concentrates in certain areas of the state space, choosing a policy can be thought of as choosing a *subset* of the environment's dynamics. In the following, a policy is represented by a subset in state space (enclosed by a directed curve) as shown in Figure 9.2.

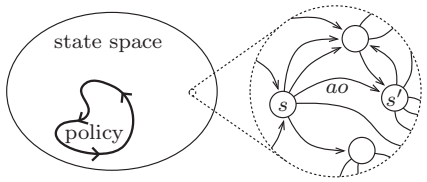



**Figure 9.2:** A Policy Diagram.

Policy diagrams are especially useful to analyze the effect of policies on different hypotheses about the environment's dynamics. For the sake of simplifying the interpretation of policy diagrams, we will assume the existence of a state space $\mathcal{T} : (\mathcal{A}\times\mathcal{O})^* \to \mathcal{S}$ mapping I/O histories into states. *Note however that no such assumptions are made to obtain the results of this section.*

### 9.4.2 Divergence Processes

The central question in this section is to investigate whether the Bayesian control rule converges to the correct control law or not. As will be obvious from the discussion in the rest of this section, this is in general not true.

As it is easily seen from (9.1), showing convergence amounts to show that the posterior distribution $P(\theta|\underline{\hat{a}o}_{<t})$ concentrates its probability mass on a subset of operation modes $\Theta_*$ having "essentially" the same output stream as $\theta_*$,

$$\sum_{\theta\in\Theta} P(a_t|\theta,\underline{ao}_{<t})P(\theta|\underline{\hat{a}o}_{<t}) \approx \sum_{\theta\in\Theta_*} P(a_t|\theta_*,\underline{ao}_{<t})P(\theta|\underline{\hat{a}o}_{<t}) \approx P(a_t|\theta_*,\underline{ao}_{<t}).$$

Hence, understanding the asymptotic behavior of the posterior probabilities

$$P(\theta|\underline{\hat{a}o}_{\leq t})$$

is crucial here. In particular, we need to understand under what conditions these quantities converge to zero. The posterior can be rewritten as

$$P(\theta|\underline{\hat{a}o}_{\leq t}) = \frac{P(\underline{\hat{a}o}_{\leq t}|\theta)P(\theta)}{\sum_{\theta'\in\Theta} P(\underline{\hat{a}o}_{\leq t}|\theta')P(\theta')} = \frac{P(\theta)\prod_{\tau=1}^{t} P(o_\tau|\theta,\underline{ao}_{<\tau}a_\tau)}{\sum_{\theta'\in\Theta} P(\theta')\prod_{\tau=1}^{t} P(o_\tau|\theta',\underline{ao}_{<\tau}a_\tau)}.$$

If all the summands but the one with index $\theta_*$ are dropped from the denominator, one obtains the bound

$$P(\theta|\underline{\hat{a}o}_{\leq t}) \leq \frac{P(\theta)}{P(\theta_*)}\prod_{\tau=1}^{t}\frac{P(o_\tau|\theta,\underline{ao}_{<\tau}a_\tau)}{P(o_\tau|\theta_*,\underline{ao}_{<\tau}a_\tau)},$$

which is valid for all $\theta_* \in \Theta$. From this inequality, it is seen that it is convenient to analyze the behavior of the following stochastic process.

**Definition 33 (Divergence Process)** The stochastic process defined by

$$d_t(\theta_*\|\theta) := \sum_{\tau=1}^{t} \ln\frac{P(o_\tau|\theta_*,\underline{ao}_{<\tau}a_\tau)}{P(o_\tau|\theta,\underline{ao}_{<\tau}a_\tau)}$$

is called the **divergence process** of $\theta$ from the reference $\theta_*$. □

Indeed, if $d_t(\theta_*\|\theta) \to \infty$ as $t \to \infty$, then

$$\lim_{t\to\infty}\frac{P(\theta)}{P(\theta_*)}\prod_{\tau=1}^{t}\frac{P(o_\tau|\theta,\underline{ao}_{<\tau}a_\tau)}{P(o_\tau|\theta_*,\underline{ao}_{<\tau}a_\tau)} = \lim_{t\to\infty}\frac{P(\theta)}{P(\theta_*)}\cdot e^{-d_t(\theta_*\|\theta)} = 0,$$

and thus clearly $P(\theta|\underline{\hat{a}o}_{\leq t}) \to 0$. Figure 9.3 illustrates simultaneous realizations of the divergence processes of a controller.

**Remark 29** Intuitively speaking, the divergence process $d_t(\theta_*\|\theta)$ is a lower bound on the "number of wrong bits predicted" by the $\theta$-th input model. □

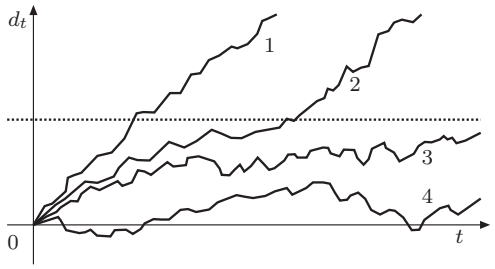



**Figure 9.3:** Realization of Divergence Processes. The divergence processes 1 to 4 are associated with a Bayesian I/O model having operation modes $\theta_1$ to $\theta_4$. The divergence processes 1 and 2 diverge, whereas 3 and 4 stay below the dotted bound. Hence, the posterior probabilities of $\theta_1$ and $\theta_2$ vanish.

**Remark 30** A divergence process is a random walk whose value at time $t$ depends on the whole history up to time $t-1$. A given divergence process can have different growth rates depending on the policy (Figure 9.4). It can happen that a divergence process stays stable under one policy, but diverges under another. □

The divergence process $d_t(\theta^* \| \theta)$ is a random variable that depends on the realization $\underline{ao}_{\leq t}$ which is drawn from

$$\prod_{\tau=1}^{t} P(a_\tau|\theta_\tau, \underline{ao}_{\leq \tau}) P(o_\tau|\theta_*, \underline{ao}_{\leq \tau} a_\tau),$$

where the $\theta_1, \theta_2, \ldots, \theta_t$ are drawn themselves from $P(\theta_1), P(\theta_2|\underline{\hat{a}o}_1), \ldots, P(\theta_t|\underline{\hat{a}o}_{<t})$.

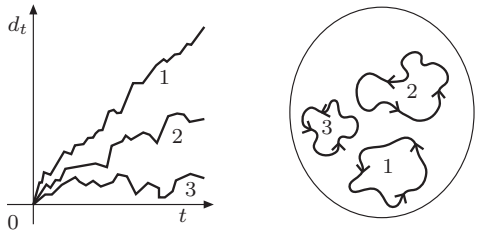



**Figure 9.4:** Policies Influence Divergence Processes. The application of different policies lead to different statistical properties of the same divergence process.

### 9.4.3 Decomposition of Divergence Processes

To deal with the heterogeneous nature of divergence processes, one introduces a temporal decomposition that demultiplexes the original process into many sub-processes *belonging to unique policies*.

**Definition 34 (Sub-Divergence)** Let $\mathcal{N}_t := \{1, 2, \ldots, t\}$ be the set of time steps up to time $t$. Let $\mathcal{T} \subset \mathcal{N}_t$, and let $\theta, \theta' \in \Theta$. Define a **sub-divergence** of $d_t(\theta_* \| \theta)$ as a random variable

$$g_{\theta'}(\theta; \mathcal{T}) := \sum_{\tau \in \mathcal{T}} \ln \frac{P(o_\tau | \theta_*, \underline{ao}_{<\tau} a_\tau)}{P(o_\tau | \theta, \underline{ao}_{<\tau} a_\tau)}$$

drawn from

$$P_{\theta'}(\{\underline{ao}_\tau\}_{\tau \in \mathcal{T}} | \{\underline{ao}_\tau\}_{\tau \in \mathcal{T}^\complement}) := \Big( \prod_{\tau \in \mathcal{T}} P(a_\tau | \theta', \underline{ao}_{<\tau}) \Big) \Big( \prod_{\tau \in \mathcal{T}} P(o_\tau | \theta_*, \underline{ao}_{<\tau} a_\tau) \Big),$$

where $\mathcal{T}^\complement := \mathcal{N}_t \setminus \mathcal{T}$ and where $\{\underline{ao}_\tau\}_{\tau \in \mathcal{T}^\complement}$ are given conditions that are kept constant.□

In this definition, $\theta'$ plays the role of the policy that is used to sample the actions in the time steps $\mathcal{T}$. Clearly, any realization of the divergence process $d_t(\theta_* \| \theta)$ can be decomposed into a sum of sub-divergences, i.e.

$$d_t(\theta_* \| \theta) = \sum_{\theta'} g_{\theta'}(\theta; \mathcal{T}_{\theta'}), \tag{9.2}$$

where $\{\mathcal{T}_\theta\}_{\theta \in \Theta}$ forms a partition of $\mathcal{N}_t$. Figure 9.5 shows an example decomposition.

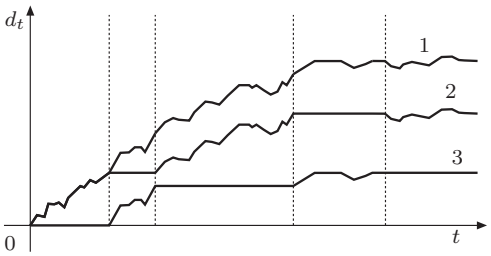



**Figure 9.5:** Decomposition of a Divergence Process (1) into Sub-Divergences (2 & 3). Note that the sub-divergences grow in disjoint time intervals.

The averages of sub-divergences will play an important rôle in the analysis. Define the average over all realizations of $g_{\theta'}(\theta; \mathcal{T})$ as

$$G_{\theta'}(\theta; \mathcal{T}) := \sum_{(\underline{ao}_\tau)_{\tau \in \mathcal{T}}} P_{\theta'}(\{\underline{ao}_\tau\}_{\tau \in \mathcal{T}} | \{\underline{ao}_\tau\}_{\tau \in \mathcal{T}^\complement})\, g_{\theta'}(\theta; \mathcal{T}).$$

Notice that for any $\tau \in \mathcal{N}_t$,

$$G_{\theta'}(\theta; \{\tau\}) = \sum_{\underline{ao}_\tau} P(a_\tau | \theta', \underline{ao}_{<\tau}) P(o_\tau | \theta_*, \underline{ao}_{<\tau} a_\tau) \ln \frac{P(o_\tau | \theta_*, \underline{ao}_{<\tau} a_\tau)}{P(o_\tau | \theta, \underline{ao}_{<\tau} a_\tau)} \geq 0,$$

because of Gibbs' inequality. In particular,

$$G_{\theta'}(\theta_*; \{\tau\}) = 0.$$

Clearly, this holds as well for any $\mathcal{T} \subset \mathcal{N}_t$:

$$\begin{aligned} \forall \theta \quad & G_{\theta'}(\theta; \mathcal{T}) \geq 0, \\ & G_{\theta'}(\theta_*; \mathcal{T}) = 0. \end{aligned} \tag{9.3}$$

### 9.4.4 Bounded Variation

In general, a divergence process is very complex: virtually all the classes of distributions that are of interest in control go well beyond the assumptions of i.i.d. and stationarity. This increased complexity can jeopardize the analytic tractability of the divergence process, such that no predictions about its asymptotic behavior can be made anymore. More specifically, if the growth rates of the divergence processes vary too much from realization to realization, then the posterior distribution over operation modes can vary qualitatively between realizations. Hence, one needs to impose a stability requirement akin to ergodicity to limit the class of possible divergence-processes to a class that is analytically tractable. For this purpose the following property is introduced.

**Definition 35 (Bounded Variation)** A divergence process $d_t(\theta_*\|\theta)$ is said to have **bounded variation** in $\Theta$ iff for any $\delta > 0$, there is a $C \geq 0$, such that for all $\theta' \in \Theta$, all $t$ and all $\mathcal{T} \subset \mathcal{N}_t$

$$\left| g_{\theta'}(\theta; \mathcal{T}) - G_{\theta'}(\theta; \mathcal{T}) \right| \leq C$$

with probability $\geq 1 - \delta$. □

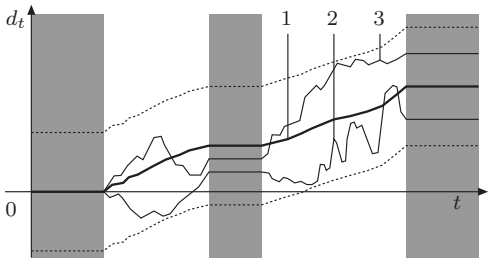



**Figure 9.6:** Bounded Variation. If a divergence process has bounded variation, then the realizations (curves 2 & 3) of a sub-divergence stay within a band around the mean (curve 1).

Figure 9.6 illustrates this property. Bounded variation is the key property that is going to be used to construct the results of this section. However, it is very restrictive. For instance, simple Bernoulli processes do not fulfill this property. The first result is that the posterior probability of the true parameter is bounded from below.

**Theorem 9 (Lower Bound of True Posterior)** *Let the set of operation modes of a controller be such that for all $\theta \in \Theta$ the divergence process $d_t(\theta_*\|\theta)$ has bounded variation. Then, for any $\delta > 0$, there is a $\lambda > 0$, such that for all $t \in \mathbb{N}$,*

$$P(\theta_* | \underline{\hat{a}o}_{\leq t}) \geq \frac{\lambda}{|\Theta|}$$

*with probability $\geq 1 - \delta$.* □

PROOF As has been pointed out in (9.2), a particular realization of the divergence process $d_t(\theta_*\|\theta)$ can be decomposed as

$$d_t(\theta_*\|\theta) = \sum_{\theta'} g_\theta(\theta'; \mathcal{T}_{\theta'}),$$

where the $g_\theta(\theta'; \mathcal{T}_{\theta'})$ are sub-divergences of $d_t(\theta_*\|\theta)$ and the $\mathcal{T}_{\theta'}$ form a partition of $\mathcal{N}_t$. However, since $d_t(\theta_*\|\theta)$ has bounded variation for all $\theta \in \Theta$, one has for all $\delta' > 0$, there is a $C(\theta) \geq 0$, such that for all $\theta' \in \Theta$, all $t \in \mathcal{N}_t$ and all $\mathcal{T} \subset \mathcal{N}_t$, the inequality

$$\left| g_\theta(\theta'; \mathcal{T}_{\theta'}) - G_\theta(\theta'; \mathcal{T}_{\theta'}) \right| \leq C(\theta)$$

holds with probability $\geq 1 - \delta'$. However, due to (9.3),

$$G_\theta(\theta'; \mathcal{T}_{\theta'}) \geq 0$$

for all $\theta' \in \Theta$. Thus,

$$g_\theta(\theta'; \mathcal{T}_{\theta'}) \geq -C(\theta).$$

If all the previous inequalities hold simultaneously then the divergence process can be bounded as well. That is, the inequality

$$d_t(\theta_*\|\theta) \geq -MC(\theta) \tag{9.4}$$

holds with probability $\geq (1 - \delta')^M$ where $M := |\Theta|$. Choose

$$\beta(\theta) := \max\{0, \ln \tfrac{P(\theta)}{P(\theta_*)}\}.$$

Since $0 \geq \ln \frac{P(\theta)}{P(\theta_*)} - \beta(\theta)$, it can be added to the right hand side of (9.4). Using the definition of $d_t(\theta_*\|\theta)$, taking the exponential and rearranging the terms one obtains

$$P(\theta_*) \prod_{\tau=1}^{t} P(o_\tau|\theta_*, \underline{ao}_{<\tau} a_\tau) \geq e^{-\alpha(\theta)} P(\theta) \prod_{\tau=1}^{t} P(o_\tau|\theta, \underline{ao}_{<\tau} a_\tau)$$

where $\alpha(\theta) := MC(\theta) + \beta(\theta) \geq 0$. Identifying the posterior probabilities of $\theta_*$ and $\theta$ by dividing both sides by the normalizing constant yields the inequality

$$P(\theta_*|\underline{\hat{a}o}_{\leq t}) \geq e^{-\alpha(\theta)} P(\theta|\underline{\hat{a}o}_{\leq t}).$$

This inequality holds simultaneously for all $\theta \in \Theta$ with probability $\geq (1 - \delta')^{M^2}$ and in particular for $\lambda := \min_\theta\{e^{-\alpha(\theta)}\}$, that is,

$$P(\theta_*|\underline{\hat{a}o}_{\leq t}) \geq \lambda P(\theta|\underline{\hat{a}o}_{\leq t}).$$

But since this is valid for any $\theta \in \Theta$, and because $\max_\theta\{P(\theta|\underline{\hat{a}o}_{\leq t})\} \geq \frac{1}{M}$, one gets

$$P(\theta_*|\underline{\hat{a}o}_{\leq t}) \geq \frac{\lambda}{M},$$

with probability $\geq 1 - \delta$ for arbitrary $\delta > 0$ related to $\delta'$ through the equation $\delta' := 1 - \sqrt[M^2]{1 - \delta}$. ∎

**Remark 31** It has been pointed by M. Hutter[4] that bounded variation can most probably be weakened to "$C$ growing sub-linearly" (which will require adapting the definitions that follow as well) and still get the convergence result of this section. □

### 9.4.5 Core

If one wants to identify the operation modes whose posterior probabilities vanish, then it is not enough to characterize them as those modes whose hypothesis does not match the true hypothesis. Figure 9.7 illustrates this problem. Here, three hypotheses along with their associated policies are shown. $H_1$ and $H_2$ share the prediction made for region $A$ but differ in region $B$. Hypothesis $H_3$ differs everywhere from the others. Assume $H_1$ is true. As long as we apply policy $P_2$, hypothesis $H_3$ will make wrong predictions and thus its divergence process will diverge as expected. However, no evidence against $H_2$ will be accumulated. It is only when one applies policy $P_1$ *for long enough time* that the agent will eventually enter region $B$ and hence accumulate counter-evidence for $H_2$.

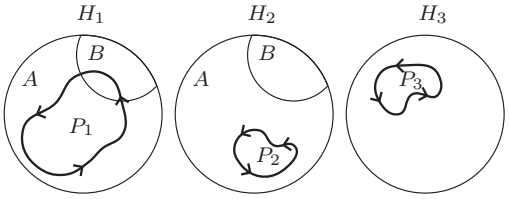



**Figure 9.7:** Problems with Disambiguation. If hypothesis $H_1$ is true and agrees with $H_2$ on region $A$, then policy $P_2$ cannot disambiguate the three hypotheses.

But what does "long enough" mean? If $P_1$ is executed only for a short period, then the controller risks not visiting the disambiguating region. But unfortunately, neither the right policy nor the right length of the period to run it are known beforehand. Hence, an agent needs a clever time-allocating strategy to test all policies for all finite time intervals. This motivates the following definition.

**Definition 36 (Core)** The **core** of an operation mode $\theta_*$, denoted as $[\theta_*]$, is the subset of $\Theta$ containing operation modes behaving like $\theta_*$ under its policy. Formally, an operation mode $\theta \notin [\theta_*]$ (i.e. is *not* in the core) iff for any $C \geq 0$, $\delta > 0$, there is a $\xi > 0$ and a $t_0 \in \mathbb{N}$, such that for all $t \geq t_0$,

$$G_{\theta_*}(\theta; \mathcal{T}) \geq C$$

with probability $\geq 1 - \delta$, where $G_{\theta_*}(\theta; \mathcal{T})$ is a sub-divergence of $d_t(\theta_* \| \theta)$, and $\mathbf{Pr}\{\tau \in \mathcal{T}\} \geq \xi$ for all $\tau \in \mathcal{N}_t$. □

In other words, if the agent was to apply $\theta_*$'s policy in each time step with probability at least $\xi$, and under this strategy the expected sub-divergence $G_{\theta_*}(\theta; \mathcal{T})$ of $d_t(\theta_* \| \theta)$ grows unboundedly, then $\theta$ is not in the core of $\theta_*$.

[4] personal communication

**Remark 32** Note that demanding a strictly positive probability of execution in each time step guarantees that the agent will run $\theta_*$ for all possible finite time-intervals. □

As the following theorem shows, the posterior probabilities of the operation modes that are not in the core vanish almost surely.

**Theorem 10 (Not in Core ⇒ Vanishing Posterior)** *Let the set of operation modes of an agent be such that for all $\theta \in \Theta$ the divergence process $d_t(\theta_* \| \theta)$ has bounded variation. If $\theta \notin [\theta_*]$, then $P(\theta | \underline{\hat{a}o}_{\leq t}) \to 0$ as $t \to \infty$ almost surely.* □

Proof The divergence process $d_t(\theta_* \| \theta)$ can be decomposed into a sum of sub-divergences (see Equation 9.2)

$$d_t(\theta_* \| \theta) = \sum_{\theta'} g_{\theta'}(\theta; \mathcal{T}_{\theta'}). \tag{9.5}$$

Furthermore, for every $\theta' \in \Theta$, one has that for all $\delta > 0$, there is a $C \geq 0$, such that for all $t \in \mathbb{N}$ and for all $\mathcal{T} \subset \mathcal{N}_t$

$$\left| g_{\theta'}(\theta; \mathcal{T}) - G_{\theta'}(\theta; \mathcal{T}) \right| \leq C(\theta)$$

with probability $\geq 1 - \delta'$. Applying this bound to the summands in (9.5) yields the lower bound

$$\sum_{\theta'} g_{\theta'}(\theta; \mathcal{T}_{\theta'}) \geq \sum_{\theta'} \big( G_{\theta'}(\theta; \mathcal{T}_{\theta'}) - C(\theta) \big)$$

which holds with probability $\geq (1 - \delta')^M$, where $M := |\Theta|$. Due to Inequality 9.3, one has that for all $\theta' \neq \theta_*$, $G_{\theta'}(\theta; \mathcal{T}_{\theta'}) \geq 0$. Hence,

$$\sum_{\theta'} \big( G_{\theta'}(\theta; \mathcal{T}_{\theta'}) - C(\theta) \big) \geq G_{\theta_*}(\theta; \mathcal{T}_{\theta_*}) - MC$$

where $C := \max_\theta \{C(\theta)\}$. The members of the set $\mathcal{T}_{\theta_*}$ are determined stochastically; more specifically, the $i$-th member is included into $\mathcal{T}_{\theta_*}$ with probability $P(\theta_* | \underline{\hat{a}o}_{\leq i}) \geq \lambda / M$ for some $\lambda > 0$ by Theorem 9. But since $\theta \notin [\theta_*]$, one has that $G_{\theta_*}(\theta; \mathcal{T}_{\theta_*}) \to \infty$ as $t \to \infty$ with probability $\geq 1 - \delta'$ for arbitrarily chosen $\delta' > 0$. This implies that

$$\lim_{t \to \infty} d_t(\theta_* \| \theta) \geq \lim_{t \to \infty} G_{\theta_*}(\theta; \mathcal{T}_{\theta_*}) - MC \nearrow \infty$$

with probability $\geq 1 - \delta$, where $\delta > 0$ is arbitrary and related to $\delta'$ as $\delta = 1 - (1 - \delta')^{M+1}$. Using this result in the upper bound for posterior probabilities yields the final result

$$0 \leq \lim_{t \to \infty} P(\theta | \underline{\hat{a}o}_{\leq t}) \leq \lim_{t \to \infty} \frac{P(\theta)}{P(\theta_*)} e^{-d_t(\theta_* \| \theta)} = 0.$$

■

### 9.4.6 Consistency

Even if an operation mode $\theta$ is in the core of $\theta_*$, i.e. given that $\theta$ is essentially indistinguishable from $\theta_*$ under $\theta_*$'s control, it can still happen that $\theta_*$ and $\theta$ have different policies. Figure 9.8 shows an example of this. The hypotheses $H_1$ and $H_2$ share region $A$ but differ in region $B$. In addition, both operation modes have their policies $P_1$ and $P_2$ respectively confined to region $A$. Note that both operation modes are in the core of each other. However, their policies are different. This means that it is unclear whether multiplexing the policies in time will ever disambiguate the two hypotheses. This is undesirable, as it could impede the convergence to the right control law.

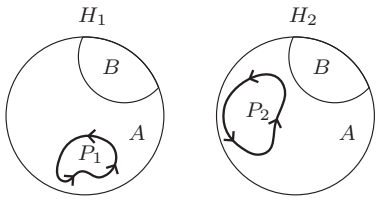



**Figure 9.8:** Inconsistent Policies. An example of inconsistent policies. Both operation modes are in the core of each other, but have different policies.

Thus, it is clear that one needs to impose further restrictions on the mapping of hypotheses into policies. With respect to Figure 9.8, one can make the following observations:

1. Both operation modes have policies that select subsets of region $A$. Therefore, the dynamics in $A$ are preferred over the dynamics in $B$.

2. Knowing that the dynamics in $A$ are preferred over the dynamics in $B$ allows us to drop region $B$ from the analysis when choosing a policy.

3. Since both hypotheses agree in region $A$, *they have to choose the same policy in order to be consistent in their selection criterion.*

This motivates the following definition.

**Definition 37 (Consistent Policies)** An operation mode $\theta$ is said to be **consistent** with $\theta_*$ iff $\theta \in [\theta_*]$ implies that for all $\varepsilon < 0$, there is a $t_0$, such that for all $t \geq t_0$ and all $\underline{ao}_{<t} a_t$,

$$\left| P(a_t|\theta, \underline{ao}_{\leq t}) - P(a_t|\theta_*, \underline{ao}_{\leq t}) \right| < \varepsilon.$$

□

In other words, if $\theta$ is in the core of $\theta_*$, then $\theta$'s policy has to converge to $\theta_*$'s policy. The following theorem shows that consistency is a sufficient condition for convergence to the right control law.

**Theorem 11 (Convergence of Bayesian Control Rule)** *Let the set of operation modes of an agent be such that: for all $\theta \in \Theta$ the divergence process $d_t(\theta_*\|\theta)$ has bounded variation; and for all $\theta, \theta_* \in \Theta$, $\theta$ is consistent with $\theta_*$. Then,*

$$P(a_t|\underline{\hat{ao}}_{<t}) \to P(a_t|\theta_*, \underline{ao}_{<t})$$

*almost surely as $t \to \infty$.* □

Proof We will use the abbreviations $p_\theta(t) := P(a_t|\theta, \underline{\hat{ao}}_{<t})$ and $w_\theta(t) := P(\theta|\underline{\hat{ao}}_{<t})$. Decompose $P(a_t|\underline{\hat{ao}}_{<t})$ as

$$P(a_t|\underline{\hat{ao}}_{<t}) = \sum_{\theta \notin [\theta_*]} p_\theta(t) w_\theta(t) + \sum_{\theta \in [\theta_*]} p_\theta(t) w_\theta(t). \tag{9.6}$$

The first sum on the right-hand side is lower-bounded by zero and upper-bounded by

$$\sum_{\theta \notin [\theta_*]} p_\theta(t) w_\theta(t) \leq \sum_{\theta \notin [\theta_*]} w_\theta(t)$$

because $p_\theta(t) \leq 1$. Due to Theorem 10, $w_\theta(t) \to 0$ as $t \to \infty$ almost surely. Given $\varepsilon' > 0$ and $\delta' > 0$, let $t_0(\theta)$ be the time such that for all $t \geq t_0(\theta)$, $w_\theta(t) < \varepsilon'$. Choosing $t_0 := \max_\theta\{t_0(\theta)\}$, the previous inequality holds for all $\theta$ and $t \geq t_0$ simultaneously with probability $\geq (1-\delta')^M$. Hence,

$$\sum_{\theta \notin [\theta_*]} p_\theta(t) w_\theta(t) \leq \sum_{\theta \notin [\theta_*]} w_\theta(t) < M\varepsilon'. \tag{9.7}$$

To bound the second sum in (9.6) one proceeds as follows. For every member $\theta \in [\theta_*]$, one has that $p_\theta(t) \to p_{\theta_*}(t)$ as $t \to \infty$. Hence, following a similar construction as above, one can choose $t_0'$ such that for all $t \geq t_0'$ and $\theta \in [\theta_*]$, the inequalities

$$\Big|p_\theta(t) - p_{\theta_*}(t)\Big| < \varepsilon'$$

hold simultaneously for the precision $\varepsilon' > 0$. Applying this to the second sum in Equation 9.6 yields the bounds

$$\sum_{\theta \in [\theta_*]} \big(p_{\theta_*}(t) - \varepsilon'\big) w_\theta(t) \leq \sum_{\theta \in [\theta_*]} p_\theta(t) w_\theta(t) \leq \sum_{\theta \in [\theta_*]} \big(p_{\theta_*}(t) + \varepsilon'\big) w_\theta(t).$$

Here $\big(p_{\theta_*}(t) \pm \varepsilon'\big)$ are multiplicative constants that can be placed in front of the sum. Note that

$$1 \geq \sum_{\theta \in [\theta_*]} w_\theta(t) = 1 - \sum_{\theta \notin [\theta_*]} w_\theta(t) > 1 - \varepsilon.$$

Use of the above inequalities allows simplifying the lower and upper bounds respectively:

$$\begin{aligned} \big(p_{\theta_*}(t) - \varepsilon'\big) \sum_{\theta \in [\theta_*]} w_\theta(t) &> p_{\theta_*}(t)(1-\varepsilon') - \varepsilon' \geq p_{\theta_*}(t) - 2\varepsilon', \\ \big(p_{\theta_*}(t) + \varepsilon'\big) \sum_{\theta \in [\theta_*]} w_\theta(t) &\leq p_{\theta_*}(t) + \varepsilon' < p_{\theta_*}(t) + 2\varepsilon'. \end{aligned} \tag{9.8}$$

Combining the inequalities (9.7) and (9.8) in (9.6) yields the final result:

$$\left| P(a_t|\underline{\hat{a}o}_{<t}) - p_{\theta_*}(t) \right| < (2+M)\varepsilon' = \varepsilon,$$

which holds with probability $\geq 1-\delta$ for arbitrary $\delta > 0$ related to $\delta'$ as $\delta' = 1 - \sqrt[M]{1-\delta}$ and arbitrary precision $\varepsilon$. ■

## 9.5 Examples

In this section we illustrate the usage of the Bayesian control rule on two examples that are very common in the reinforcement learning literature: multi-armed bandits and Markov decision processes. As a reminder, a summary of the Bayesian control rule is given in Table 9.1.

**Bayesian Control Rule:** Given a set of operation modes $\{P(\cdot|\theta,\cdot)\}_{\theta\in\Theta}$ over interaction sequences in $\mathcal{Z}^\infty$ and a prior distribution $P(\theta)$ over the parameters $\Theta$, the probability of the action $a_{t+1}$ is given by

$$P(a_{t+1}|\underline{\hat{a}o}_{\leq t}) = \sum_\theta P(a_{t+1}|\theta, \underline{ao}_{\leq t}) P(\theta|\underline{\hat{a}o}_{\leq t}), \tag{9.9}$$

where the posterior probability over operation modes is given by the recursion

$$P(\theta|\underline{\hat{a}o}_{\leq t}) = \frac{P(o_t|\theta, \underline{ao}_{<t}) P(\theta|\underline{\hat{a}o}_{<t})}{\sum_{\theta'} P(o_t|\theta', \underline{ao}_{<t}) P(\theta'|\underline{\hat{a}o}_{<t})}.$$

**Table 9.1:** Summary of the Bayesian control rule.

### 9.5.1 Bandit Problems

Consider the *multi-armed bandit problem* Robbins (1952). The problem is stated as follows. Suppose there is an $N$-armed bandit, i.e. a slot-machine with $N$ levers. When pulled, lever $i$ provides a reward drawn from a Bernoulli distribution with a bias $h_i$ specific to that lever. That is, a reward $r = 1$ is obtained with probability $h_i$ and a reward $r = 0$ with probability $1 - h_i$. The objective of the game is to maximize the time-averaged reward through iterative pulls. There is a continuum range of stationary strategies, each one parameterized by $N$ probabilities $\{s_i\}_{i=1}^N$ indicating the probabilities of pulling each lever. The difficulty arising in the bandit problem is to balance reward maximization based on the knowledge already acquired with attempting new actions to further improve knowledge. This dilemma is known as the exploration versus exploitation tradeoff Sutton and Barto (1998).

This is an ideal task for the Bayesian control rule, because each possible bandit has a known optimal agent. Indeed, a bandit can be represented by an $N$-dimensional bias vector $\theta = [\theta_1, \ldots, \theta_N] \in \Theta = [0;1]^N$. Given such a bandit, the optimal policy consists in pulling the lever with the highest bias. That is, an operation mode is given by:

$$h_i = P(o_t = 1|\theta, a_t = i) = \theta_i \qquad s_i = P(a_t = i|\theta) = \begin{cases} 1 & \text{if } i = \max_j\{\theta_j\}, \\ 0 & \text{else.} \end{cases}$$

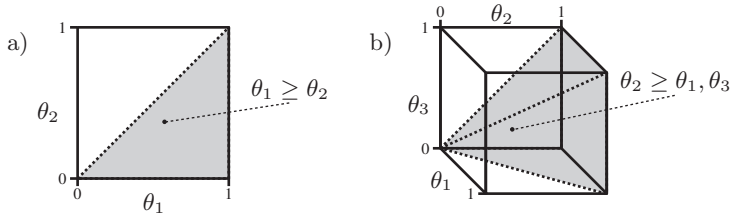



**Figure 9.9:** The space of bandit configurations can be partitioned into $N$ regions according to the optimal lever. Panel a and b show the 2-armed and 3-armed bandit cases respectively.

To apply the Bayesian control rule, it is necessary to fix a prior distribution over the bandit configurations. Assuming a uniform distribution, the Bayesian control rule is

$$P(a_{t+1} = i|\underline{\hat{a}o}_{\leq t}) = \int_\Theta P(a_{t+1} = i|\theta) P(\theta|\underline{\hat{a}o}_{\leq t}) \tag{9.10}$$

with the update rule given by

$$P(\theta|\underline{\hat{a}o}_{\leq t}) = \frac{P(\theta)\prod_{\tau=1}^t P(o_\tau|\theta, a_\tau)}{\int_\Theta P(\theta')\prod_{\tau=1}^t P(o_\tau|\theta', a_\tau)\, d\theta'} = \prod_{j=1}^N \frac{\theta_j^{r_j}(1-\theta_j)^{f_j}}{\mathrm{B}(r_j+1, f_j+1)} \tag{9.11}$$

where $r_j$ and $f_j$ are the counts of the number of times a reward has been obtained from pulling lever $j$ and the number of times no reward was obtained respectively. Observe that here the summation over discrete operation modes has been replaced by an integral over the continuous space of configurations. In the last expression we see that the posterior distribution over the lever biases is given by a product of $N$ Beta distributions. Thus, sampling an action amounts to first sample an operation mode $\theta$ by obtaining each bias $\theta_i$ from a Beta distribution with parameters $r_i + 1$ and $f_i + 1$, and then choosing the action corresponding to the highest bias $a = \arg\max_i \theta_i$. The pseudo-code can be seen in Algorithm 1.

**Simulation:** The Bayesian control rule described above has been compared against two other agents: an $\varepsilon$-greedy strategy with decay (on-line) and Gittins indices (off-line). The test bed consisted of bandits with $N = 10$ levers whose biases were drawn uniformly at the beginning of each run. Every agent had to play 1000 runs for 1000

**Algorithm 1:** BCR bandit.

**for** $i = 1, \ldots, N$ **do**
  Initialize $r_i$ and $f_i$ to zero.
*Main cycle:* **for** $t = 1, 2, 3, \ldots$ **do**
  Sample $m$ using (9.11).
  *Interaction:*
  Set $a \leftarrow \arg\max_i m_i$ and issue $a$.
  Obtain $o$ from environment.
  *Update belief:*
  **if** $o = 1$ **then**
    $r_a = r_a + 1$
  **else**
    $f_a = f_a + 1$

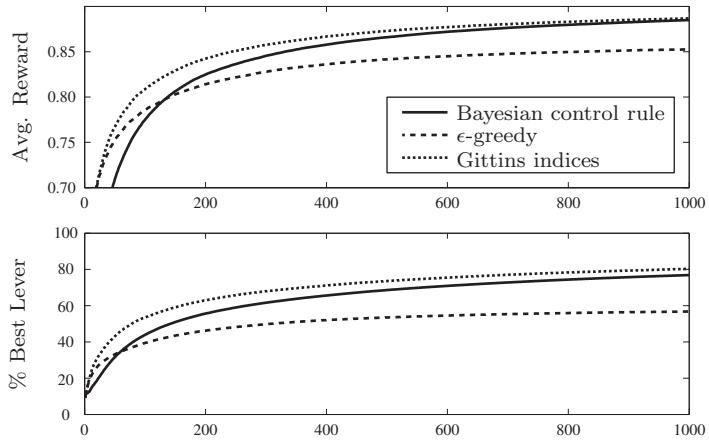



**Figure 9.10:** Comparison in the $N$-armed bandit problem of the Bayesian control rule (solid line), an $\varepsilon$-greedy agent (dashed line) and using Gittins indices (dotted line). 1,000 runs have been averaged. The top panel shows the evolution of the average reward. The bottom panel shows the evolution of the percentage of times the best lever was pulled.

time steps each. Then, the performance curves of the individual runs were averaged. The $\varepsilon$-greedy strategy selects a random action with a small probability given by $\varepsilon\alpha^{-t}$ and otherwise plays the lever with highest expected reward. The parameters have been determined empirically to the values $\varepsilon = 0.1$, and $\alpha = 0.99$ after several test runs. They have been adjusted in a way to maximize the average performance in the last trials of our simulations. For the Gittins method, all the indices were computed up to horizon 1300 using a geometric discounting of $\alpha = 0.999$, i.e. close to one to approximate the time-averaged reward. The results are shown in Figure 9.10.

It is seen that $\varepsilon$-greedy strategy quickly reaches an acceptable level of performance, but then seems to stall at a significantly suboptimal level, pulling the optimal lever only 60% of the time. In contrast, both the Gittins strategy and the Bayesian control rule show essentially the same asymptotic performance, but differ in the initial transient phase where the Gittins strategy significantly outperforms the Bayesian control rule. There are at least three observations that are worth making here. First, Gittins indices have to be pre-computed off-line. The time complexity scales quadratically with the horizon, and the computations for the horizon of 1300 steps took several hours on our machines. In contrast, the Bayesian control rule could be applied without pre-computation. Second, even though the Gittins method actively issues the optimal information gathering actions while the Bayesian control rule passively samples the actions from the posterior distribution over operation modes, in the end both methods rely on the convergence of the underlying Bayesian estimator. This implies that both methods have the same information bottleneck, since the Bayesian estimator requires the same amount of information to converge. Thus, active information gathering actions only affect the utility of the transient phase, not the permanent state. Other efficient algorithms for bandit problems can be found in the literature (Auer, CesaBianchi, and Fischer, 2002).

### 9.5.2 Markov Decision Processes

A Markov Decision Process (*MDP*) is defined as a tuple $(\mathcal{X}, \mathcal{A}, T, r)$: $\mathcal{X}$ is the state space; $\mathcal{A}$ is the action space; $T_a(x; x') = \mathbf{Pr}(x'|a, x)$ is the probability that an action $a \in \mathcal{A}$ taken in state $x \in \mathcal{X}$ will lead to state $x' \in \mathcal{X}$; and $r(x, a) \in \mathcal{R} := \mathbb{R}$ is the immediate reward obtained in state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$. The interaction proceeds in time steps $t = 1, 2, \ldots$ where at time $t$, action $a_t \in \mathcal{A}$ is issued in state $x_{t-1} \in \mathcal{X}$, leading to a reward $r_t = r(x_{t-1}, a_t)$ and a new state $x_t$ that starts the next time step $t + 1$. A stationary closed-loop control policy $\pi : \mathcal{X} \to \mathcal{A}$ assigns an action to each state. For MDPs there always exists an optimal stationary deterministic policy and thus one only needs to consider such policies. In undiscounted MDPs the average reward per time step for a fixed policy $\pi$ with initial state $x$ is defined as $\rho^\pi(x) = \lim_{t\to\infty} \mathbf{E}^\pi[\frac{1}{t}\sum_{\tau=0}^{t} r_\tau]$. It can be shown Bertsekas (1987) that $\rho^\pi(x) = \rho^\pi(x')$ for all $x, x' \in \mathcal{X}$ under the assumption that the Markov chain for policy $\pi$ is ergodic. Here, we assume that the MDPs are ergodic for all stationary policies.

In order to keep the intervention model particularly simple[5], we follow the Q-notation of Watkins (1989). The optimal policy $\pi^*$ can then be characterized in terms of the optimal average reward $\rho$ and the optimal relative Q-values $Q(x,a)$ for each state-action pair $(x,a)$ that are solutions to the following system of non-linear equations (Singh, 1994): for any state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$,

$$\begin{aligned} Q(x,a) + \rho &= r(x,a) + \sum_{x' \in \mathcal{X}} \mathbf{Pr}(x'|x,a) \Big[ \max_{a'} Q(x',a') \Big] \\ &= r(x,a) + \mathbf{E}_{x'} \Big[ \max_{a'} Q(x',a') \Big| x, a \Big]. \end{aligned} \tag{9.12}$$

The optimal policy can then be defined as $\pi^*(x) := \arg\max_a Q(x,a)$ for any state $x \in \mathcal{X}$.

Again this setup allows for a straightforward solution with the Bayesian control rule, because each learnable MDP (characterized by the Q-values and the average reward) has a known solution $\pi^*$. Accordingly, an operation mode $\theta$ is given by $\theta = [Q, \rho] \in \Theta = \mathbb{R}^{|\mathcal{A}| \times |\mathcal{O}| + 1}$. To obtain a likelihood model for inference over $\theta$, we realize that Equation 9.12 can be rewritten such that it predicts the instantaneous reward $r(x,a)$ as the sum of a mean instantaneous reward $\xi_\theta$ plus a noise term $\nu$ given the $Q$-values and the average reward $\rho$ for the MDP labeled by $\theta$

$$r(x,a) = \underbrace{Q(x,a) + \rho - \max_{a'} Q(x',a')}_{\text{mean instantaneous reward } \xi_\theta(x,a,x')} \underbrace{+ \max_{a'} Q(x',a') - \mathbf{E}[\max_{a'} Q(x',a') | x, a]}_{\text{noise } \nu}$$

Assuming that $\nu$ can be reasonably approximated by a normal distribution $\mathrm{N}(0, 1/p)$ with precision $p$, we can write down a likelihood model for the immediate reward $r$ using the Q-values and the average reward, i.e.

$$P(r|\theta, x, a, x') = \sqrt{\frac{p}{2\pi}} \exp\Big\{ -\frac{p}{2} (r - \xi_\theta(x,a,x'))^2 \Big\}. \tag{9.13}$$

In order to determine the intervention model for each operation mode, we can simply exploit the above properties of the $Q$-values, which gives

$$P(a|\theta, x) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q(x,a') \\ 0 & \text{else.} \end{cases} \tag{9.14}$$

To apply the Bayesian control rule, the posterior distribution $P(\theta|\hat{a}_{\leq t}, x_{\leq t})$ needs to be computed. Fortunately, due to the simplicity of the likelihood model, one can

[5]The "brute-force" adaptive agent for this problem would roughly look as follows. First, the agent starts with a prior distribution over all MDPs, e.g. product of Dirichlet distributions over the transition probabilities. Then, in each cycle, the agent samples a full transition matrix from the distribution and solves it using dynamic programming. Once it has computed the optimal policy, it uses it to issue the next action, and then discards the policy. Subsequently, it updates the distribution over MDPs using the next observed state. However, in the main text we follow a different approach that avoids solving an MDP in every time step.

easily devise a conjugate prior distribution and apply standard inference methods (see derivation in Section 9.8 at the end of this Chapter). Actions are again determined by sampling operation modes from this posterior and executing the action suggested by the corresponding intervention models. The resulting algorithm is very similar to Bayesian Q-learning Dearden, Friedman, and Russell (1998); Dearden, Friedman, and Andre (1999), but differs in the way actions are selected. The pseudo-code is listed in Algorithm 2.

**Algorithm 2:** BCR-MDP Gibbs sampler.

Initialize entries of $\lambda$ and $\mu$ to zero.
Set initial state to $x \leftarrow x_0$.
**for** $t = 1, 2, 3, \ldots$ **do**
  *Gibbs sweep:*
  Sample $\rho$ using (9.19).
  **for** $Q(y, b)$ *of visited states* **do**
    Sample $Q(y, b)$ using (9.20).
  *Interaction:*
  Set $a \leftarrow \arg\max_{a'} Q(x, a')$ and issue $a$.
  Obtain $o = (r, x')$ from environment.
  *Update hyperparameters:*
  $\mu(x, a, x') \leftarrow \frac{\lambda(x,a,x')\mu(x,a,x') + p\, r}{\lambda(x,a,x') + p}$
  $\lambda(x, a, x') \leftarrow \lambda(x, a, x') + p$
  Set $x \leftarrow x'$.

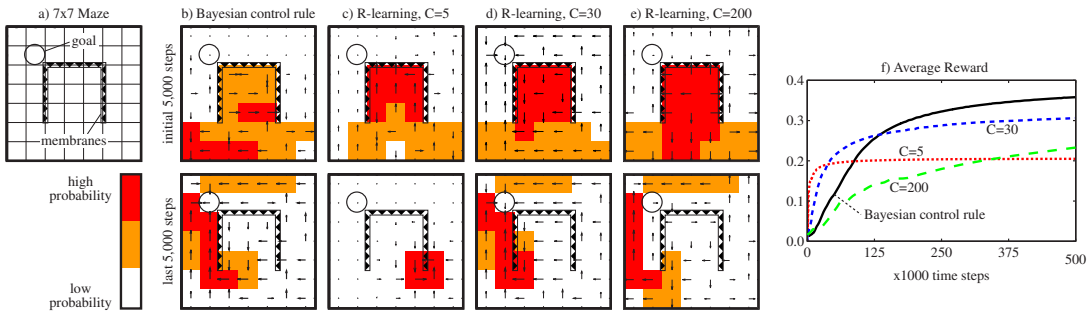



**Figure 9.11:** Results for the $7 \times 7$ grid-world domain. Panel (a) illustrates the setup. Columns (b)-(e) illustrate the behavioral statistics of the algorithms. The upper and lower row have been calculated over the first and last 5,000 time steps of randomly chosen runs. The probability of being in a state is color-encoded, and the arrows represent the most frequent actions taken by the agents. Panel (f) presents the curves obtained by averaging ten runs.

**Simulation:** We have tested our MDP-agent in a grid-world example. To give an intuition of the achieved performance, the results are contrasted with those achieved by R-learning. We have used the R-learning variant presented in the work of Singh (1994, Algorithm 3) together with the uncertainty exploration strategy Mahadevan (1996). The corresponding update equations are

$$\begin{aligned} Q(x,a) &\leftarrow (1-\alpha)Q(x,a) + \alpha\big(r - \rho + \max_{a'} Q(x',a')\big) \\ \rho &\leftarrow (1-\beta)\rho + \beta\big(r + \max_{a'} Q(x',a') - Q(x,a)\big), \end{aligned} \tag{9.15}$$

where $\alpha, \beta > 0$ are learning rates. The exploration strategy chooses with fixed probability $p_{\text{exp}} > 0$ the action $a$ that maximizes $Q(x,a) + \frac{C}{F(x,a)}$, where $C$ is a constant, and $F(x,a)$ represents the number of times that action $a$ has been tried in state $x$. Thus, higher values of $C$ enforce increased exploration.

In a study Mahadevan (1996), a grid-world is described that is especially useful as a test bed for the analysis of RL algorithms. For our purposes, it is of particular interest because it is easy to design experiments containing *suboptimal limit-cycles*. Figure 9.11, panel (a), illustrates the $7 \times 7$ grid-world. A controller has to learn a policy that leads it from any initial location to the goal state. At each step, the agent can move to any adjacent space (up, down, left or right). If the agent reaches the goal state then its next position is randomly set to any square of the grid (with uniform probability) to start another trial. There are also "one-way membranes" that allow the agent to move into one direction but not into the other. In these experiments, these membranes form "inverted cups" that the agent can enter from any side but can only leave through the bottom, playing the role of a local maximum. Transitions are stochastic: the agent moves to the correct square with probability $p = \frac{9}{10}$ and to any of the free adjacent spaces (uniform distribution) with probability $1 - p = \frac{1}{10}$. Rewards are assigned as follows. The default reward is $r = 0$. If the agent traverses a membrane it obtains a reward of $r = 1$. Reaching the goal state assigns $r = 2.5$. The parameters chosen for this simulation were the following. For our MDP-agent, we have chosen hyperparameters $\mu_0 = 1$ and $\lambda_0 = 1$ and precision $p = 1$. For R-learning, we have chosen learning rates $\alpha = 0.5$ and $\beta = 0.001$, and the exploration constant has been set to $C = 5$, $C = 30$ and to $C = 200$. A total of 10 runs were carried out for each algorithm. The results are presented in Figure 9.11 and Table 9.2. R-learning only learns the optimal policy given sufficient exploration (panels d & e, bottom row), whereas the Bayesian control rule learns the policy successfully. In Figure 9.11f, the learning curve of R-learning for $C = 5$ and $C = 30$ is initially steeper than the Bayesian controller. However, the latter attains a higher average reward around time step 125,000 onwards. We attribute this shallow initial transient to the phase where the distribution over the operation modes is flat, which is also reflected by the initially random exploratory behavior.

| | Average Reward |
|---|---|
| BCR | $0.3582 \pm 0.0038$ |
| R-learning, $C = 200$ | $0.2314 \pm 0.0024$ |
| R-learning, $C = 30$ | $0.3056 \pm 0.0063$ |
| R-learning, $C = 5$ | $0.2049 \pm 0.0012$ |

**Table 9.2:** Average reward attained by the different algorithms at the end of the run. The mean and the standard deviation has been calculated based on 10 runs.

## 9.6 Critical Issues

**Problems of Bayesian methods.** The Bayesian control rule treats an adaptive control problem as a Bayesian inference problem. Hence, all the problems typically associated with Bayesian methods carry over to agents constructed with the Bayesian control rule. These problems are of both analytical and computational nature. For example, there are many probabilistic models where the posterior distribution does not have a closed-form solution. Also, exact probabilistic inference is in general computationally very intensive. Even though there is a large literature in efficient/approximate inference algorithms for particular problem classes Bishop (2006), not many of them are suitable for on-line probabilistic inference in more realistic environment classes.

**Bayesian control rule versus Bayes-optimal control.** Directly maximizing the (subjective) expected utility for a given environment class is not the same as minimizing the expected relative entropy for a given class of operation modes. *The two methods are based on different assumptions and optimality principles.* As such, the Bayesian control rule is not a Bayes-optimal controller. Indeed, it is easy to design experiments where the Bayesian control rule converges exponentially slower (or does not converge at all) than a Bayes-optimal controller to the maximum utility. Consider the following simple example: Environment 1 is a $k$-state MDP in which only $k$ consecutive actions $A$ reach a state with reward $+1$. Any interception with a $B$-action leads back to the initial state. Consider a second environment which is like the first but actions $A$ and $B$ are interchanged. A Bayes-optimal controller figures out the true environment in $k$ actions (either $k$ consecutive $A$'s or $B$'s). Consider now the Bayesian control rule: The optimal action in Environment 1 is $A$, in Environment 2 is $B$. A uniform $(\frac{1}{2}, \frac{1}{2})$ prior over the operation modes stays a uniform posterior as long as no reward has been observed. Hence the Bayesian control rule chooses at each time-step $A$ and $B$ with equal probability. With this policy it takes about $2^k$ actions to accidentally choose a row of $A$'s (or $B$'s) of length $k$. From then on the Bayesian control rule is optimal too. So a Bayes-optimal controller converges in time $k$, while the Bayesian control rule needs exponentially longer. One way to remedy this problem might be to allow the Bayesian control rule to sample actions from the same operation mode for several time steps in a row rather than randomizing controllers in every cycle. However, if one considers non-stationary environments this strategy can also break down. Consider, for example,

an increasing MDP with $k = \lceil 10\sqrt{t} \rceil$, in which a Bayes-optimal controller converges in 100 steps, while the Bayesian control rule does not converge at all in most realizations, because the boundedness assumption is violated.

## 9.7 Relation to Existing Approaches

Some of the ideas underlying this work are not unique to the Bayesian control rule. The following is a selection of previously published work in the recent Bayesian reinforcement learning literature where related ideas can be found.

**Compression principles.** In the literature, there is an important amount of work relating compression to intelligence (MacKay, 2003; Hutter, 2004a). In particular, it has been even proposed that compression ratio is an objective quantitative measure of intelligence (Mahoney, 1999). Compression has also been used as a basis for a theory of curiosity, creativity and beauty (Schmidhuber, 2009).

**Mixture of experts.** Passive sequence prediction by mixing experts has been studied extensively in the literature (Cesa-Bianchi and Lugosi, 2006). In a study on online-predictors (Hutter, 2004b), Bayes-optimal predictors are mixed. Bayes-mixtures can also be used for universal prediction (Hutter, 2003). For the control case, the idea of using mixtures of expert-controllers has been previously evoked in models like the MOSAIC-architecture (Haruno, Wolpert, and Kawato, 2001). Universal learning with Bayes mixtures of experts in reactive environments has been studied in the work of Poland and Hutter (2005) and Hutter (2002).

**Stochastic action selection.** The idea of using actions as random variables, and the problems that this entails, has been expressed in the work of Hutter (2004a, Problem 5.1). The study in this chapter can be regarded as a thorough investigation of this open problem. Other stochastic action selection approaches are found in the thesis of Wyatt (1997) who examines exploration strategies for (PO)MDPs, in learning automata (Narendra and Thathachar, 1974) and in probability matching (Duda, Hart, and Stork, 2001) amongst others. In particular, the thesis discusses theoretical properties of an extension to *probability matching* in the context of multi-armed bandit problems. There, it is proposed to choose a lever according to how likely it is to be optimal and it is shown that this strategy converges, thus providing a simple method for guiding exploration.

## 9.8 Derivation of Gibbs Sampler for MDP Agent

Inserting the likelihood given in Equation 9.13 into Equation 9.9 of the Bayesian control rule, one obtains the following expression for the posterior

$$\begin{aligned} P(\theta|\hat{a}_{\leq t}, o_{\leq t}) &= \frac{P(x'|\theta,x,a)P(r|\theta,x,a,x')P(\theta|\hat{a}_{<t},o_{<t})}{\int_\Theta P(x'|\theta',x,a)P(r|\theta',x,a,x')P(\theta'|\hat{a}_{<t},o_{<t})\,d\theta'} \\ &= \frac{P(r|\theta,x,a,x')P(\theta|\hat{a}_{<t},o_{<t})}{\int_\Theta P(r|\theta',x,a,x')P(\theta'|\hat{a}_{<t},o_{<t})\,d\theta'}, \end{aligned} \tag{9.16}$$

where we have replaced the sum by an integration over $\theta'$, the finite-dimensional real space containing only the average reward and the Q-values of the observed states, and where we have simplified the term $P(x'|\theta,x,a)$ because it is constant for all $\theta' \in \Theta$.

The likelihood model $P(r|\theta',x,a,x')$ in Equation 9.16 encodes a set of independent normal distributions over the immediate reward with means $\xi_\theta(x,a,x')$ indexed by triples $(x,a,x') \in \mathcal{X}\times\mathcal{A}\times\mathcal{X}$. In other words, given $(x,a,x')$, the rewards are drawn from a normal distribution with unknown mean $\xi_\theta(x,a,x')$ and known variance $\sigma^2$. The sufficient statistics are given by $n(x,a,x')$, the number of times that the transition $x \to x'$ under action $a$, and $\bar{r}(x,a,x')$, the mean of the rewards obtained in the same transition. The conjugate prior distribution is well known and given by a normal distribution with hyperparameters $\mu_0$ and $\lambda_0$:

$$P(\xi_\theta(x,a,x')) = \mathrm{N}(\mu_0, 1/\lambda_0) = \sqrt{\frac{\lambda_0}{2\pi}}\exp\Big\{-\tfrac{\lambda_0}{2}\big(\xi_\theta(x,a,x')-\mu_0\big)^2\Big\}. \tag{9.17}$$

The posterior distribution is given by

$$P(\xi_\theta(x,a,x')|\hat{a}_{\leq t}, o_{\leq t}) = \mathrm{N}(\mu(x,a,x'), 1/\lambda(x,a,x'))$$

where the posterior hyperparameters are computed as

$$\begin{aligned} \mu(x,a,x') &= \frac{\lambda_0\,\mu_0 + p\,n(x,a,x')\,\bar{r}(x,a,x')}{\lambda_0 + p\,n(x,a,x')} \\ \lambda(x,a,x') &= \lambda_0 + p\,n(x,a,x'). \end{aligned} \tag{9.18}$$

By introducing the shorthand $V(x) := \max_a Q(x,a)$, we can write the posterior distribution over $\rho$ as

$$P(\rho|\hat{a}_{\leq t}, o_{\leq t}) = \mathrm{N}(\bar{\rho}, 1/S) \tag{9.19}$$

where

$$\bar{\rho} = \frac{1}{S}\sum_{x,a,x'} \lambda(x,a,x')(\mu(x,a,x') - Q(x,a) + V(x')),$$

$$S = \sum_{x,a,x'} \lambda(x,a,x').$$

The posterior distribution over the Q-values is more difficult to obtain, because each $Q(x,a)$ enters the posterior distribution both linearly and non-linearly through $\mu$. However, if we fix $Q(x,a)$ within the max operations, which amounts to treating each

$V(x)$ as a constant within a single Gibbs step, then the conditional distribution can be approximated by

$$P(Q(x,a)|\hat{a}_{\leq t}, o_{\leq t}) \approx \mathrm{N}\Big(\bar{Q}(x,a), 1/S(x,a)\Big) \tag{9.20}$$

where

$$\bar{Q}(x,a) = \frac{1}{S(x,a)} \sum_{x'} \lambda(x,a,x')(\mu(x,a,x') - \rho + V(x')),$$
$$S(x,a) = \sum_{x'} \lambda(x,a,x').$$

We expect this approximation to hold because the resulting update rule constitutes a contraction operation that forms the basis of most stochastic approximation algorithms Mahadevan (1996). As a result, the Gibbs sampler draws all the values from normal distributions. In each cycle of the adaptive controller, one can carry out several Gibbs sweeps to obtain a sample of $\theta$ to improve the mixing of the Markov chain. However, our experimental results have shown that a *single Gibbs sweep per state transition* performs reasonably well. Once a new parameter vector $\theta$ is drawn, the Bayesian control rule proceeds by taking the optimal action given by Equation 9.14. Note that only the $\mu$ and $\lambda$ entries of the transitions that have occurred need to be represented explicitly; similarly, only the Q-values of visited states need to be represented explicitly.

## 9.9 Historical Remarks & References

The Bayesian control rule has been entirely developed by the author and D. A. Braun and it has been first published in Ortega and Braun (2010c). The convergence proof was published later in Ortega and Braun (2010a). See also Braun and Ortega (2010). The name "Bayesian control rule" has been suggested by Z. Ghahramani.

Some of the ideas underlying this work are not unique to the Bayesian control rule. The following is a selection of previously published work in the recent Bayesian control literature where related ideas can be found.

*Compression principles.* In the literature, there is an important amount of work relating compression to intelligence (MacKay, 2003; Hutter, 2004a). In particular, it has been even proposed that compression ratio is an objective quantitative measure of intelligence (Mahoney, 1999). Compression has also been used as a basis for a theory of curiosity, creativity and beauty (Schmidhuber, 2009).

*Mixture of experts.* Passive sequence prediction by mixing experts has been studied extensively in the literature (Cesa-Bianchi and Lugosi, 2006). In a study on online-predictors (Hutter, 2004b), Bayes-optimal predictors are mixed. Bayes-mixtures can also be used for universal prediction (Hutter, 2003). For the control case, the idea of using mixtures of expert-controllers has been previously evoked in models like the MOSAIC-architecture (Haruno et al., 2001). Universal learning with Bayes mixtures of experts in reactive environments has been studied in the work of Poland and Hutter (2005) and Hutter (2002).

*Stochastic action selection.* The idea of using actions as random variables, and the problems that this entails, has been expressed in the work of Hutter (2004a, Problem 5.1). The present

chapter can be regarded as a thorough investigation of this open problem. Other stochastic action selection approaches are found in the thesis of Wyatt (1997) who examines exploration strategies for (PO)MDPs, in learning automata (Narendra and Thathachar, 1974) and in probability matching (Duda et al., 2001) amongst others. In particular, the thesis discusses theoretical properties of an extension to *probability matching* in the context of multi-armed bandit problems. There, it is proposed to choose a lever according to how likely it is to be optimal and it is shown that this strategy converges, thus providing a simple method for guiding exploration.

*Relative entropy criterion.* The usage of a minimum relative entropy criterion to derive control laws underlies the KL-control methods developed in the work of Todorov (2006, 2009) and Kappen et al. (2009). There, it has been shown that a large class of optimal control problems can be solved very efficiently if the problem statement is reformulated as the minimization of the deviation of the dynamics of a controlled system from the uncontrolled system. A related idea is to conceptualize planning as an inference problem (Toussaint, Harmeling, and Storkey, 2006). This approach is based on an equivalence between maximization of the expected future return and likelihood maximization which is both applicable to MDPs and POMDPs. Algorithms based on this duality have become an active field of current research. See for example the work of Rasmussen and Deisenroth (2008), where very fast model-based reinforcement learning techniques are used for control in continuous state and action spaces.

# Chapter 10

# Discussion

Despite of recent major theoretical achievements in the field of artificial intelligence, the current state-of-the-art implementations are still far away from even displaying insect intelligence. Given this situation, the two main questions addressed in this thesis are:

1. Are there limitations imposed by the mathematical foundations of classical agency?
2. If yes, how do we formulate new foundations that overcome these limitations?

## 10.1 Summary

This thesis has been organized in two parts, summarized as follows.

1. The first part contains a concise presentation of the foundations of classical agency: namely the formalizations of decision making and learning. The first includes: SEU theory, the framework of decision making under uncertainty; the maximum SEU principle to choose the optimal solution; and its application to the design of autonomous systems, culminating in the Bellman optimality equations. The second includes: Bayesian probability theory, the theory for reasoning under uncertainty that extends logic; and Bayes-Optimal agents, the application of Bayesian probability theory to the design of optimal adaptive agents.
2. Then, two major problems of the maximum SEU principle are highlighted: the prohibitive computational costs and the need for the causal precedence of the choice of the policy.
3. The second part tackles the two aforementioned problems. First, an information-theoretic notion of resources in autonomous agents is established. Second, a framework for resource-bounded agency is introduced. This includes: a maximum bounded SEU principle that is derived from a set of axioms of utility; an axiomatic model of probabilistic causality, which is applied for the formalization of autonomous systems having uncertainty over their policy and environment; and the Bayesian control rule, derived from the maximum bounded SEU principle and

the model of causality, implementing a stochastic adaptive control law that deals with the case where autonomous agents are uncertain about their policy and environment.

## 10.2 What are the contributions?

**Bounded SEU.** There is abundant literature tackling the problem of bounded rationality, whose purpose is: (a) capturing aspects of human decision making that contradicts rationality; or (b) understanding how resource costs affect decision making. This thesis introduces a related view whose main goal is to present an axiomatic formalization of bounded rationality that encompasses classical rationality as a limit case when resource costs vanish. More specifically, the utility-information conversion factor $\alpha$ controls the tradeoff between information and utility, and the energy-information conversion factor $\gamma$ controls the tradeoff between information and energy. Real autonomous systems have $\alpha > 0$, and hence, under the view of the presented framework, perfect rationality is only an idealization. Conceptually, the distinguishing feature of bounded SEU (compared to classical SEU) is that it provides an explanatory framework for approximations.

**Causality.** Causality is a field that has historically been highly controversial, and it has been only recently that mathematical formalizations have started to find wider acceptance. Still, so far these formalizations have not clarified the connection to measure theory, the mathematically rigorous theory of probability. The framework introduced in this thesis does a step towards this direction. Simultaneously, this thesis clarifies the importance of causal consideration in agent designs. More specifically, the Bayesian I/O model (Section 8.3.1), which is really a causal model, allows enriching the classical Bayesian autonomous system, having uncertainty only over its environment, with having uncertainty over its very policy.

**Bayesian control rule.** Agents that are constructed following the maximum SEU principle have to carry out massive computations *before* they even have had a single interaction with its environment. To bypass this limitation, most practical algorithms implement autonomous systems that "discover" their policy *during* the interactions with the environment. In this thesis, we have used the framework of bounded SEU and causal reasoning to derive the Bayesian control rule: a rule that allows constructing a natural class adaptive agents having uncertainty over their policies. Formally, the *Bayesian control rule* for outputs is the probabilistic equivalent to the *predictive distribution* for inputs. Furthermore, this thesis presents a convergence proof for the Bayesian control rule under a very restricted setting.

## 10.3 What is missing?

**Understanding the Implications of the Relations.** If one is willing to accept the connections between decision theory, information theory and thermodynamics that this thesis puts forward, then one obtains a potentially very fertile ground for novel ideas and reinterpretations. While this thesis shows some of the benefits of adopting this unified view, it also leaves many questions unanswered. For instance, the utility-information conversion factor $\alpha$ controls the cost of translating resources into utilities. On a very abstract level, one could blame the failure of approximations to the very large value of $\alpha$. But what does $\alpha$ mean in practice? How can it be influenced? Examples such as this abound and need to be addressed in the future.

**Descriptive Power.** While the bounded SEU framework introduced in this thesis has a theoretical appeal due to its properties and simplicity, it remains to be seen whether it can explain human decision making, and especially, whether it can explain the experimental evidence (Allais, 1953; Ellsberg, 1961; Kahneman and Tversky, 1979; Machina, 1987; Kreps, 1988) that contradicts perfect rationality. In particular, it would be important to verify whether the causality framework and/or the Bayesian control rule can predict aspects of human decision making.

**Intelligence Measure.** Legg and Hutter (2006) proposed a formal measure of machine intelligence. While this measure is a synthesis of many informal definitions of human intelligence that have been given by experts, it has been constructed mainly borrowing ideas from the theory of universal artificial intelligence, staying thus within the paradigm of agency with unlimited resources. It would be interesting to test whether this intelligence measure can be accommodated to include agency with bounded resources.

**Game Theory.** It is not at all clear how the design of autonomous systems connects to the game theoretic literature. The fundamental difference lies in the assumptions. In artificial intelligence and control theory, one assumes a dynamical model of the environment first, and then constructs a suitable agent. Meanwhile, in game theory, one instead *only* assumes a utility function describing the environment's preferences. This assumption, however, does not provide enough information to derive the dynamical model of the environment, since this model must depend on the assumptions the environment makes about the agent. Under this point of view, additional solution concepts (collectively called **equilibria**) other than the maximum expected utility principle are required to derive the resulting behavior of the interaction system. In the context of this thesis, an important point to be verified is to investigate whether the Bayesian I/O model for agents is "essentially" equivalent to the Bayesian game framework (Harsanyi, 1967–1968), and whether game theoretic concepts can be extended to the case of resource-bounded agency.

# References


M. Allais. Le comportment de l'homme rationnel devant la risque: critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21:503–546, 1953.

Dana Angluin. Computational learning theory: survey and selected bibliography. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, STOC '92, pages 351–369, New York, NY, USA, 1992.

F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, mar 1963.

R. B. Ash. *Information Theory*. New York: Interscience, 1965.

P. Auer, N. CesaBianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

R. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 1763.

P. Beame. A general sequential time-space tradeoff for finding unique elements. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, STOC '89, pages 197–203, New York, NY, USA, 1989. ACM.

P. Beame, T. S. Jayram, and M. Saks. Time-space tradeoffs for branching programs. *J. Comput. Syst. Sci.*, 63:542–572, December 2001.

P. Beame, M. Saks, X. Sun, and E. Vee. Time-space trade-off lower bounds for randomized computation of decision problems. *J. ACM*, 50:154–195, March 2003.

R. E. Bellman. Dynamic programming, 1957.

C. H. Bennett. Logical reversibility of computation. *IBM Journal of Research and Development*, 17(6):525–532, 1973.

C. H. Bennett. The thermodynamics of computationa review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.

D. Bernoulli. Specimen theoriae novae de mensara sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 1738. (trans. in 1954, *Econometrica*).

## REFERENCES


D. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Upper Saddle River, NJ, 1987.

P. Billingsley. *Ergodic Theory and Information*. R. E. Krieger Pub. Co., 1978.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

K. Borch. A note on uncertainty and indifference curves. *The Review of Economic Studies*, 36(1):1–4, 1969.

E. Borel. *Leçons sur la théorie des fonctions*. Gauthier-Villars, Paris, 1898.

A. Borodin and S. Cook. A time-space tradeoff for sorting on a general sequential model of computation. In *Proceedings of the twelfth annual ACM symposium on Theory of computing*, STOC '80, pages 294–301, New York, NY, USA, 1980. ACM.

D. A. Braun and P. A. Ortega. A minimum relative entropy principle for adaptive control in linear quadratic regulators. In *Proceedings of the 7th international conference on informatics in control, automation and robotics*, page (in press), 2010.

H. J. Bremermann. Quantum noise and information. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, California, 1965. Univ. of California Press.

L. Brillouin. Maxwell's demon cannot operate: Information and entropy i. *Journal of Applied Physics*, 22:334–337, 1951.

L. Brillouin. *Science and Information Theory*. New York: Academic Press., 1956.

H. B. Callen. *Thermodynamics and an Introduction to Themostatistics*. John Wiley & Sons, 2nd edition, 1985.

C. F. Camerer and M. Weber. Recent developments in modelling preferences: uncertainty and ambiguity. *J. Risk Uncertain.*, 5:325–370, 1992.

J. C. Candeal, J. R. De Miguel, E. Induráin, and G. B. Mehta. Utility and entropy. *Economic Theory*, 17:233–238, 2001.

N. Cartwright. *Nature's Capacities and Their Measurement*. Claredon Press, Oxnard, 1989.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

T. M. Cover and J. A. Thomas. *Elements of information theory*. New York: Wiley-Interscience, 1st edition, 1991.

R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins, 1961.

A. P. Dawid. Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London, 2007.

B. De Finetti. La prévision: Ses lois logiques, ses sources subjectives. In *Annales de l'Institut Henri Poincaré*, volume 7, pages 1–68, 1937.

R. Dearden, N. Friedman, and S. Russell. Bayesian q-learning. In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 761–768, Menlo Park, CA, US, 1998. American Association for Artificial Intelligence.

R. Dearden, N. Friedman, and D. Andre. Model based bayesian exploration. In *In Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–159, 1999.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley & Sons, Inc., second edition, 2001.

M. O'G. Duff. *Optimal learning: computational procedures for bayes-adaptive markov decision processes*. PhD thesis, University of Massachusetts—Amherst, 2002. Director-Andrew Barto.

E. Eells. *Probabilistic Causality*. Cambridge University Press, 1991.

D. Ellsberg. Risk, ambiguity and the savage axioms. *The Quaterly Journal of Economics*, 75:643–669, 1961.

R. P. Feynman. *Feynman Lectures on Computation*. Westview Press, 2nd edition, 2000.

P. C. Fishburn. *The Foundations of Expected Utility*. D. Reidel Publishing, Dordrecht, 1982.

R. A. Fisher. *Statistical Methods for Research Workers*. Macmillan Pub. Co., 13th edition, 1970.

M. Fréchet. Définition de l'intégrale d'une fonctionnelle étendue à un ensemble abstrait. *Comptes Rendus de l'Académie des sciences*, 1915.

D. Gabor. Light and information. *Progress in Optics*, 1:111–153, 1964.

R. Gallager. *Information Theory and Reliable Communication*. New York: John Wiley and Sons, 1968.

H. Goldstein. *Classical Mechanics*. Addison-Wesley, 2nd edition, 1980.

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

J. C. Harsanyi. Games with incomplete information played by "bayesian" players, i–iii. *Management Science*, 14:159–182, 320–334, 486–502, 1967–1968.

## REFERENCES


M. Haruno, D. M. Wolpert, and M. Kawato. Mosaic model for sensorimotor learning and control. *Neural Computation*, 13:2201–2220, 2001.

M. Hutter. Self-optimizing and pareto-optimal policies in general environments based on bayes-mixtures. In *COLT*, 2002.

M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–997, 2003.

M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004a.

M. Hutter. Online prediction – bayes versus experts. Technical report, July 2004b. Presented at the EU PASCAL Workshop on Learning Theoretic and Bayesian Inductive Principles (LTBIP-2004).

E. T. Jaynes and Larry G. Bretthorst. *Probability Theory: The Logic of Science: Books*. Cambridge University Press, 2003.

H. Jeffreys. *The Theory of Probability*. The Clarendon Press, 1939.

M. I. Jordan, Z. Ghahramani, and L. K. Saul. Hidden Markov decision trees. In *Advances in Neural Information Processing Systems 9*, 1997.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

B. Kappen, V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. *arXiv:0901.0633*, 2009.

M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

G. Keller. *Equilibrium States in Ergodic Theory*. London Mathematical Society Student Texts. Cambridge University Press, 1998.

A. I. Khinchin. *Mathematical Foundations of Information Theory*. New York: Dover, 1957.

F. H. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, Boston, 1921.

A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

A. N. Kolmogorov. *Three approaches to the quantitative definition of information*. International Journal of Computer Mathematics, 1968.

L. G. Kraft. *A Device for Quantizing, Grouping, and Coding Amplitude Modulated Pulses*. Ms thesis, Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, MA, 1949.

D. M. Kreps. *Notes on the Theory of Choice*. Westview Press, 1988.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951.

R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.

P. S. Laplace. Mémoires sur la probabilité des causes par les évènements. *Mémoires de mathématique et des physiques presenteés à l'Académie royale des sciences, par divers savans, & lûs dans ses assemblées*, 6:621–656, 1774.

H. Lebesgue. *Leçons sur l'intégration et la recherche des fonctions primitives.* 1904.

S. Legg. *Machine Super Intelligence*. PhD thesis, Department of Informatics, University of Lugano, June 2008.

S. Legg and M. Hutter. A formal measure of machine intelligence. In *Annual machine learning conference of Belgium and The Netherlands (Benelearn-2006)*, Ghent, 2006.

D. Lewis. *Counterfactuals*. Cambridge, MA: Harvard University Press, 1973.

M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications (Texts in Computer Science)*. Springer, February 2008.

R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley, 1959.

M. Machina. Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, 1:121–154, 1987.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

S. Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1-3):159–195, 1996.

M. V. Mahoney. Text compression as a test for artificial intelligence. In *AAAI/IAAI*, pages 486–502, 1999.

J. C. Maxwell. Letter to P. G. Tait, 11 December 1867, 1867.

W. S. McCulloch and W. Pitts. A logical calculus of ideas immanent in neural activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

## REFERENCES


B. McMillan. Two inequalities implied by unique decipherability. *IEEE Trans. Information Theory*, 2(4):115–116, 1956.

D. H. Mellor. *The Facts of Causation*. Routledge, 1995.

D. Michie. Game-playing and game-learning automata. *Advances in Programming & Non-Numerical Computation*, pages 183–200, 1966.

K. Narendra and M. A. L. Thathachar. Learning automata - a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-4(4):323–334, July 1974.

J. Neyman. *First course in probability and statistics*. Holt, New York, 1950.

N. J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann Publishers, San Francisco, 1998.

R. Nozick. Newcomb's problem and two principles of choice. In N. Rescher, editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. Reidel, 1969.

P. A. Ortega and D. A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010a.

P. A. Ortega and D. A. Braun. A conversion between utility and information. In *The third conference on artificial general intelligence*, pages 115–120, Paris, 2010b. Atlantis Press.

P. A. Ortega and D. A. Braun. A bayesian rule for adaptive control based on causal interventions. In *The third conference on artificial general intelligence*, pages 121–126, Paris, 2010c. Atlantis Press.

P. A. Ortega and D. A. Braun. An axiomatic formalization of bounded rationality based on a utility-information equivalence. *arXiv:1007.0940*, 2010d.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1999.

C. H. Papadimitriou. *Computational Complexity*. Adison Wesley, 1993.

D. C. Parkes. Bounded rationality. Technical report, University of Pennsylvania, 1997.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

J. Poland and M. Hutter. Defensive universal learning with experts. In *ALT*, 2005.

K. R. Popper. *The Logic of Scientific Discovery (Routledge Classics)*. Routledge, 1934.

T. W. Pratt and M. V. Zelkowitz. *Programming Languages: Design and Implementation*. Prentice Hall, 4th edition, 2000.

F. P. Ramsey. *The Foundations of Mathematics and Other Logical Essays*, chapter 'Truth and Probability'. Harcourt, Brace and Co., 1926. posthumously published in 1931.

C. E. Rasmussen and M. P. Deisenroth. *Recent Advances in Reinforcement Learning*, volume 5323 of *Lecture Notes on Computer Science, LNAI*, chapter Probabilistic Inference for Fast Learning in Control, pages 229–242. Springer-Verlag, 2008.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Societry*, 58:527–535, 1952.

A. Rosenblueth, N. Wiener, and J. Bigelow. Behavior, purpose, and teleology. *Philosophy of Science*, 10:18–24, 1943.

A. Rubinstein. *Modeling Bounded Rationality*. MIT Press, 1988.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition edition, 2009.

S. Russell and E. Wefald. Principles of metareasoning. *Artificial Intelligence*, 49:361–395, 1991.

J. E. Savage. *Models of Computation: Exploring the Power of Computing*. Addison Wesley Publishing Company, 1998.

L. J. Savage. *The Foundations of Statistics*. John Wiley and Sons, New York, 1954.

J. Schmidhuber. Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE*, 48(1):21–32, 2009.

G. Shafer. *The Art of Causal Conjecture*. MIT Press, 1996.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, Jul and Oct 1948.

H. Simon. *Models of Bounded Rationality*. MIT Press, 1982.

H. A. Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69:99–188, 1955.

S. P. Singh. Reinforcement learning algorithms for average-payoff markovian decision processes. In *National Conference on Artificial Intelligence*, pages 700–705, 1994.

M. Sipser. *Introduction to the Theory of Computation*. PWS Pub. Co., 1996.

R. J. Solomonoff. A formal theory of inductive inference. part 1 & 2. *Information and Control*, 7:1–22, 224–254, 1964.

## REFERENCES


P. Spirtes and R. Scheines. *Causation, Prediction, and Search, Second Edition*. MIT Press, 2001.

R. Stalnacker. *Ifs: Conditionals, Belief, Decision, Chance, and Time*, chapter A Theory of Conditionals., pages 41–56. Dordrecht: Reidel, 1968.

P. Suppes. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company, 1970.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

L. Szilard. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Zeitschrift für Physik*, 53:840–856, 1929.

E. Todorov. Linearly solvable markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, pages 1369–1376, 2006.

E. Todorov. General duality between optimal control and estimation. In *Proceedings of the 47th conference on decision and control*, pages 4286–4292, 2008.

E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences U.S.A.*, 106:11478–11483, 2009.

M. Toussaint, S. Harmeling, and A. Storkey. Probabilistic inference for solving (po)mdps, 2006.

M. Tribus and E. C. McIrvine. Energy and information. *Scientific American*, 225: 179–188, 1971.

J. N. Tsitsiklis. Computational complexity in Markov decision theory. *HERMIS—An International Journal of Computer Mathematics and its Applications*, 9(1):45–54, 2007.

J. Veness, K. S. Ng, M. Hutter, and D. Silver. Reinforcement learning via aixi approximation. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

J. Veness, K. S. Ng, M. Hutter, Uther W., and D. Silver. A monte-carlo aixi approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011.

J. Venn. *The Logic of Chance*. Macmillan Pub. Co., London, 1st edition, 1866.

R. von Mises. Grundlagen der wahrscheinlichkeitsrechnung. *Mathemat. Zeitsch.*, 5: 52–99, 1919.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

C. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, Cambridge, England, 1989.

N. Wiener. *Cybernetics, Second Edition: or the Control and Communication in the Animal and the Machine*. The MIT Press, 1965.

J. Wyatt. *Exploration and Inference in Learning from Reinforcement*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1997.

# Index