Informational Constraints and Organisation of Behaviour

Sander G. van Dijk

University of Hertfordshire

A thesis submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of *Doctor of Philosophy*

July, 2013

Abstract

Based on the view of an agent as an information processing system, and the premise that for such a system it is evolutionary advantageous to be parsimonious with respect to informational burden, an information-theoretical framework is set up to study behaviour under information minimisation pressures. This framework is based on the existing method of relevant information, which is adopted and adapted to the study of a range of cognitive aspects.

Firstly, the model of a simple reactive actor is extended to include layered decision making and a minimal memory, in which it is shown that these aspects can decrease some form of bandwidth requirements in an agent, but at the cost of an increase at a different stage or moment in time, or for the system as a whole. However, when combined, they *do* make it possible to operate with smaller bandwidths at each part of the cognitive system, *without* increasing the bandwidth of the whole or lowering performance.

These results motivate the development of the concept of *look-ahead information*, which extends the relevant information method to include time, and future informational effects of immediate actions in a more principled way. It is shown that this concept can give rise to intrinsic drives to avoid uncertainty, simplify the environment, and develop a predictive memory.

Next, the framework is extended to incorporate a set of goals, rather than deal with just a single task. This introduces the task description as a new source of relevant information, and with that the concept of *relevant goal information.* Studying this quantity results in several observations: minimising goal information bandwidth results in ritualised behaviour; relevant goal and state information may to some point be exchanged for one another without affecting the agent's performance; the dynamics of goal information give rise to a natural notion of sub-goals; bottlenecks on goal memory, and a measure of efficiency on the use of these bottlenecks, provide natural abstractions of the environment, and a global reference frame that supersedes local features of the environment.

Finally, it is shown how an agent or species could actually arrive at having a large repertoire of goals and accompanying optimal sensors and behaviour, while under a strong information-minimisation pressure. This is done by introducing an informational model of sensory evolution, which indicates that a fundamental information-theoretical law may underpin an important evolutionary catalyst; namely, even a fully minimal sensor can carry additional information, dubbed here *concomitant information*, that is required to unlock

the actual relevant information, which enables a minimal agent to still explore, enter and acquire different niches, accelerating a possible evolution to higher acuity and behavioural abilities.

Acknowledgments

> “Knowledge is in the end based on acknowledgement.”
>
> *Ludwig Wittgenstein*

Many people have been instrumental in arriving at the end product that you have in front of you, to whom I wish to express the greatest of gratitude.

Firstly, I am very thankful to Dr. Daniel Polani, my principal supervisor, who convinced me to come to Hertfordshire and embark on a wonderful journey through the forest of information theory. His extensive knowledge, interesting discussions, and instructive guidance while giving substantial space for own exploration give a student little more to wish for. Many thanks also goes to Prof. Chrystopher Nehaniv, my secondary supervisor, and the rest of the staff at the school of Computer Science for their support.

I was honoured to be part of the SEPIAs, whose information ninja skills never let me down, and who formed the best research group I could have hoped to enter, including some of the most intelligent, interesting, and inspiring people I know. Thanks for bringing life to Hatfield! Of course the same thanks go out to the other members of the Adaptive Systems Research Group, with whom the best of times were had, and who made the last years a great and memorable experience.

Furthermore, I would like to thank the University of Hertfordshire for their extensive support in my RoboCup efforts, which has made it possible to set aside time to vent the stress of high velocity research into this leisureful hobby of mine. Without this support I would have missed out on many life changing adventures, opportunities, and on meeting many wonderful people from around the world. Meeting these people, learning from them, and working together with them has instilled an unimaginable amount of knowledge and experience upon me. I am very grateful to all the members of the Bold Hearts throughout the years who I have enjoyed these adventures with, and eternal thanks goes to the original Little Green BATS, who have started it all; without them I and this work would not have been here.

Finally, I thank my parents, Wybe and Olga van Dijk, for supporting me in all I do, for their lessons, and for ensuring I had each option opened for me to push myself as much as possible.

But most importantly, my whole heart goes out to my Kinga, who has made this journey so much brighter, who carried at least the weight I did along the way, and who I love so much. Nagyon szeretlek.

Contents

Chapter 1

Introduction

> “Real understanding will only come from distillation of general principles at a higher level, [...] if we look forward far enough into the future, we are driven to seek general principles rather than detailed minutiae.”
>
> *Richard Dawkins*

1.1 Foundation: Life and Trade-Offs

Life is full of trade-offs: should we buy a car, or invest the money? Should I pursue a specific goal, or keep my options open? Can you achieve more by thinking less?

In general, an organism may be seen as the result of a possibly large set of trade-offs, which are induced by different and opposing evolutionary pressures. For example, a predator may be driven to become bigger and stronger to enable it to overpower larger prey, while at the same time there may be a pressure towards lighter and leaner bodies, such that it can better outrun its meal. One can expect that species under evolutionary pressure are driven towards the optimal trade-off, in this case where it would be as fast as it can be, while still keeping enough strength to overpower a prey.

Similarly, there may be yet another pressure to be more observing or cunning, so there is less need to be strong or fast. Larger brains, and larger or more precise sensors to supply such brains with more detailed input, open up a wider range of behaviour for an organism, possibly enhancing its performance in its niche. However, this comes at a significant energetic cost, as exemplified by multiple studies; e.g. the eye of a resting fly accounts for 10% of its energy consumption [57], and about 20% of human energy consumption is accounted for by our brain [44].

Such numbers suggest that there is indeed a real trade-off to be made under evolutionary pressure. This assumed process is schematically visualised in Fig. 1.1: with growing cognitive capacities, and the burden associated with employing them, the achievable performance of an organism is expected to grow monotonically, tracing out some optimal trade-off curve. The existence of such a trade-off curve induces two possible, complementary questions: 1) what is the maximum performance that can be achieved under a given limit on cognitive burden, and 2) what is the minimal cognitive burden that must be employed to achieve a given performance level?

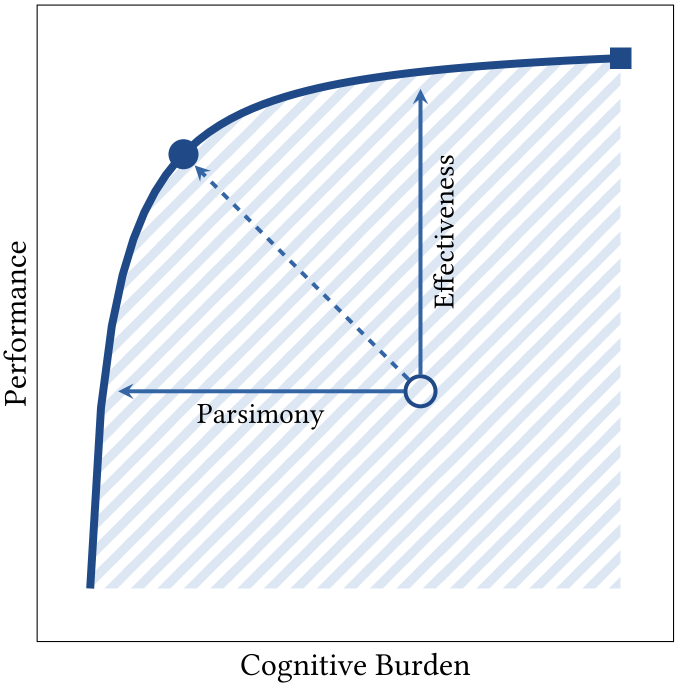


Figure 1.1: Trade-off between cognitive burden and behavioural performance. The available cognitive power restricts the range of feasible behavioural performance, denoted by the shaded area. The boundary of this area (solid line) traces the *optimal trade-off curve*, i.e. the highest performance achievable without surpassing a given load, or, equivalently, the minimal load needed to achieve a given level of fitness, with the global optimum with the highest performance at the tip (square). A species below this curve will feel evolutionary pressures to be cognitively more efficient, and/or use its cognitive power more parsimoniously (solid arrows), moving it towards a point on the optimal curve (dotted arrow, filled circle).

A species that operates underneath this curve can be said to behave sub-optimally, as it spends more energy on its cognitive processes than necessary to achieve its current fitness. From this I arrive at the following two dual hypotheses, that form the foundation of the work in this thesis:

Hypothesis 1. The Effectiveness Hypothesis: *A species or agent is under evolutionary pressure to increase fitness through more effective utilisation of any superfluous cognitive capacity.*

Hypothesis 2. The Parsimony Hypothesis: *A species or agent is under evolutionary pressure to degenerate sensory and cognitive capacity, to be more efficient and do away with unneeded energy consumption.*

These pressures effect a move towards the optimal curve, until it is hit.

There are some arguments that one could put forward as possible opposition against the two hypotheses that form the foundation of this thesis as described above, which I would like to get out of the way directly.

'Redundancy Reduction' was Wrong

Forms of the Parsimony Hypothesis have been around for a long time, and the idea that it can be studied with information theory came about shortly after Shannon's creation of the field. Notable proponents were Barlow and Attneave [12, 6], who predicted that the brain would minimise the amount of redundancy of information that it processes. However, armed with decades of empirical results and new knowledge, this prediction is now discredited. In the words of Barlow [11]:

> “Originally both Attneave and I strongly emphasised the economy that could be achieved by re-coding sensory messages to take advantage of their redundancy, but two points have become clear since those early days. First, anatomical evidence shows that there are very many more neurons at higher levels in the brain, suggesting that redundancy does not decrease, but actually increases. Second, the obvious forms of compressed, non-redundant, representation would not be at all suitable for the kinds of task that brains have to perform with the information represented; [...].”

How can an abundance of redundancy still fit with the idea of parsimony? The main answer lies in what Barlow points out that redundancy reduction is *not*: stripping away information that is unimportant, which in contrast *is* the type of parsimony that I will study here. My hypothesis is that an agent is driven to structure its cognitive system and/or its actions such that it can discard as much unimportant information as possible. The observation that the information that *is* important is represented with a lot of redundancy, that this redundancy increases at higher levels, and the idea that this redundancy *must* be available for a brain to be able to perform its tasks only increases this drive: the cognitive cost of operating under the optimal curve is exacerbated by this redundancy.

There is no Trade-Off in Nature

Experiments have shown that rods in the retina of toads can respond already to just a single photon [15]. A viper's pit heat sensor can react to heat differences of 0.003° [23], and is directionally accurate enough to enable a blindfolded viper to strike a target within five degrees of dead centre [70]. Finally, it was shown that the inner ear detects forces comparable to the thermal-noise limit [33], and the fly *Ormia ochracea* can detect a time difference of $1.5\mu s$ between auditory impulses arriving at its ears, and use this to localise sound sources within 2 degrees azimuth [65].

From these examples, it seems that evolution is simply seeking out the far end of the trade-off curve between sensory acuity, and the cognitive burden that comes with processing the high amount of information they give. Is there then actually an optimal trade-off curve, on which each point is locally optimal? Or is there just a global optimum that nature is moving towards and that is only limited by the laws of physics? In other words, is the only interesting point of Fig. 1.1 the upper right corner?

I would argue this is not the case; evolution has no predefined direction, and is unmoved by our amazement at the hyperacuity that is achieved. There is a plethora of examples where evolution drove the degeneration of vestigialized characters, not excluding sensory and cognitive apparatus, such as in blind cavefish [42] and a range of other cave dwelling animals, flightless birds, loss of hearing in moths, and countless more [37].

So, it seems that the earlier examples only show *one* possible outcome, where there has been enough pressure to develop highly accurate sensors and specialised neural facility to work with these, to outweigh the costs of operating them, but this is not necessarily the final outcome. The occurrence of organisms operating on or close to the physical limits however still poses an interesting question: how can this end of the trade-off curve actually be achieved? It may seem unlikely, as the pressures visualised in Fig. 1.1 only effect a drive 'backwards' towards the optimal curve. In this view it seems difficult to climb the curve towards the end. I will investigate this problem in more detail in Chap. 7.

1.2 Behaviour, Cognition, Sensing and Acting

In the previous section some key concepts were mentioned such as 'behaviour' and 'cognition', which are some of those concepts that are hard to define in a general yet specific way, and satisfactory for all purposes. Moreover, they involve more of these seemingly abstract terms, such as 'sensing' and 'acting'. One may have an 'I know it when I see it' idea of these concepts: when you see or feel something, you are sensing; when you pick it up and manipulate it, you are acting; and when you are solving that difficult puzzle, or try to figure out the local public transport network, there is certainly some cognition going on. But how would you formulate concrete definitions that capture it all? Is acting just what you do with your hands, or is any muscle movement an action? Is speaking? And what if you talk to yourself? Does cognition consist just of those obviously mentally taxing thought processes, or is cognition also applied in those mundane every day, semi-conscious, or even unconscious behaviours, such as cycling, making coffee, or breathing?

To be able to develop this thesis, I must make these concepts more explicit, and put in place a concrete model. Purposefully, these will still be as abstract as possible: the goal

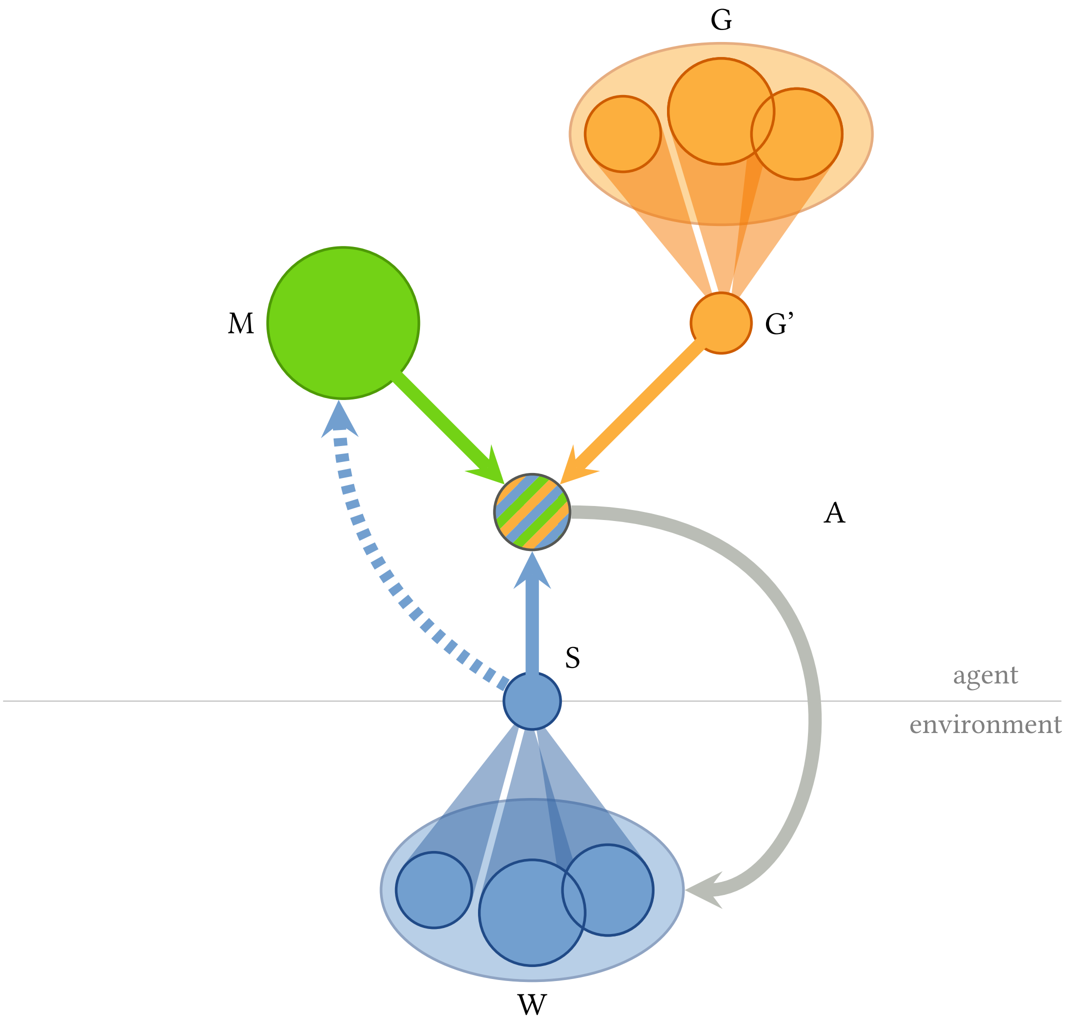


Figure 1.2: Sketch of the full cognitive model studied in this thesis. An agent interacts with the world W through its sensors S and actions A. It may have a memory M, to retain information that can guide future actions, and internal drives that induce behaviour directed towards some goals G. The currently relevant knowledge about these goals can be made accessible for concrete action decisions in a goal working memory G'.

is to present properties and fundamental phenomena of cognition and behaviour without needing to specify, and thereby limit, the physical systems that could underlie the studied processes. With that in mind, Fig. 1.2 shows the model of behaviour generation that I will use. It is very much in line with classic definitions of agents and their so called *Perception-Action loop* [80], and consists of the following parts:

Agent/World The model makes a hard distinction between two interacting entities: the agent and the world. Although this may seem a valid distinction, it is important to note that skips over some possible issues. For instance, if you rely on artificial prostheses, do they belong to you as an agent, or to the world? More generally, some argue that any item used actively to support cognition can be considered an extension of the mind, instead of as part of the environment [29]. Also in cases of tight symbiosis such as lichen, or 'swarm intelligence' such as with ants or bees, one may say that the full cognitive process is not contained with a single individual organism. Finally, in artificial systems wholly different and novel interactions may be possible that may blur the line: consider a group of robots partly controlled by a central computer, or robots that can join up and separate by linking and disconnecting their bodies and circuitry at will. For the purpose of this thesis, I contend that the place of the boundary depends on the observer, and make no assumption on where the distinction is made, only on the fact that this distinction can be made, and that the world and agent can be described concretely and separately.

Sensing With the distinction between the agent and the world in place, sensing is then the reception of any information from the world by the agent, or any change in the agent directly caused by interaction with the world, which affects its cognitive process and/or its actions.

Acting Acting covers the remainder of interaction with the world: any manipulation of the state of the world directly by the agent.

Cognition In this model, cognition consists of whatever process that links sensing with acting. This includes some really basic links for which the term 'cognition' as used in daily language may seem an exaggeration, such as a pain reflex that makes you lift your hand from a hot stove before you are aware of it. However, this definition allows one to study a system in terms of information processing requirements and make fundamental statements about it without concern of the implementation: the reflex may actually be a result of lightning fast mental processing of all possible outcomes of not lifting your hand (although I hope the work in this thesis will support a bias towards a simpler assumption), it still processed the same input information with the same outcome. This is not to say that one is unable to break down the process and study the possible role of its parts. This I will do for 2 major components: memory and internal drives. The first is modeled purely as a sensory memory, because by definition sensations capture all external information on which actions are based. The second is introduced to drive those differences in behaviour that are not readily explained by current sensations and memory, with traditional examples such as hunger and thirst resulting in different behavior in the kitchen.

Behaviour Finally, behaviour encompass everything, as the combined result of the sensing, cognitive, and acting processes over time.

1.3 Research Questions and Contributions

In this thesis, I will study the trade-off between cognition and behaviour, as well as some purely cognitive trade-offs, based on the hypotheses stated in Sec. 1.1. Specifically, I will study the following questions:

- What trade-offs are involved in sensing, cognition and performance?
- How can such trade-offs be rigorously formalised?
- Under such a formalisation, how does one find optimal trade-offs?
- How do different assumptions on the cognitive architecture determine different trade-offs?
- What fundamental properties of cognitive processes and behaviour can be studied, determined, and predicted from such a formalisation and its optimal solutions?

These investigations result in the following novel contributions contained in this thesis:

Methods

1. Extensions to the application of the relevant information method (c.q. Sec. 2.6) to a range of cognitive models, including models of distributed decision making, memory, and goal-directedness, and to different or combined information constraints in these models.
2. The novel concept and associated methods of look-ahead information, which finds informational trade-offs that take into account future informational effects of their actions.
3. A learning method, based on look-ahead information, as the first demonstration of an agent acquiring an information minimising policy solely based on experience, in an on-line, situated scenario.
4. A method to study information transitions in the working memory of a goal-directed agent through sampling.
5. A framework to study drives in sensor evolution from an informational standpoint.

Contributions to Knowledge

1. Quantitative evidence that a properly organised multi-component cognitive structure can alleviate cognitive burden, as compared to a single unit.
2. Quantitative results that show that an agent that can take into account future informational effects of actions and that is under a drive to limit sensory bandwidth:

- is driven to avoid future states with unpredictable outcome
- is driven to make its world simpler
- can develop a memory that is useful in predicting future state, even when it is not needed to increase performance

3. That it is possible for different forms of information in an agent's decision making process to be substituted for each other, with no effect on performance.

4. That natural sub-goals and segmentations, while in other research often imposed on the scenario, arise intrinsically and naturally from the viewpoint of information minimisation.

5. That the fundamental properties of a perception-action loop, arising from information-theoretic laws, induces a strong candidate to be an evolutionary catalyst, in the form of *concomitant information*: additional information not relevant in itself, but required to access information that is relevant.

1.4 Scope

Again, concepts such as behaviour and cognition are very broad, and there are a vast amount of interesting research questions and as many approaches to treat them. To remain manageable, the research in this thesis will be contained to the behaviour and cognition of a single agent, using computational models of elemental discrete action selection scenarios; the goal of this work is to offer a formal framework in which to study the aforementioned hypotheses, with presupposing, and depending on, as few details of agents and their environments as possible, while still allowing the observation of a range of important phenomena. The scenarios and experiments are chosen with the aim to strike this balance. Their relevance will further be discussed in the final chapter, where I will also discuss the prospects and opportunities of moving beyond these scenarios.

The basic hypotheses are inspired by nature, and the framework that will be developed is aimed to capture aspects of biological life. Throughout the thesis I will relate results to findings in nature wherever possible. However, I will not claim that the methods used to study different phenomena map directly to specific biological processes, but rather have as purpose to analyse the properties and outcomes of such processes. For instance, I will not explain what specific evolutionary steps are required to arrive at a sensory channel with minimum capacity, but rather what general inherent structures can be observed once such a channel is achieved.

Similarly, I will also at various points relate my results to work on artificial intelligence and machine learning, on which a large part of the framework is built, but there the aim is to provide a novel, in a sense more fundamental viewpoint to regard this work from. None of the methods used are aimed to be practical ways to achieve intelligent artificial behavior, although I will discuss how the results may point towards natural concepts that may guide its development.

1.5 Outline

The remainder of this thesis is organised as follows. The next chapter provides an overview of the technical background of the work presented later. In it I will give an introduction to familiarise the reader with the formal frameworks and methods, and introduce some vocabulary that underpin this work, as well as provide some examples and interpretations of their results. After that, Chap. 3 will review other work that is related to these methods and the investigations performed in this thesis. The four chapters following these overviews present the original work performed.

The first of these, Chap. 4, will apply the parsimony hypothesis, and extend the relevant information methods to cognitive systems with more intricate spatial and temporal organisation, and show how such differences in organisation affect informational requirements. The results of this chapter will form a guideline for the remainder. In Chap. 5 I will present the novel concept of look-ahead information, and with it develop new methods to find and study cognitive trade-offs that take into account future informational effects. Chapter 6 applies the relevant information method to agents that have internal drives that induce different goal-directed behaviour at different times. Finally, Chap. 7 will formulate sensor evolution in terms of the informational framework, and show how a strong evolutionary catalysing effect can be grounded in fundamental informational laws.

Together, Chapters 4–7 study parts, and combinations of parts, of the cognitive model sketched in Sec. 1.2 and Fig. 1.2. More precisely, they study the effects of the parsimony hypothesis, or constraints, on the bandwidth of the information channels in this cognitive model.

Chapter 8 rounds up the results of these chapters in a general discussion. Finally, App. A provides some more technical details of the methods discussed throughout this thesis.

Chapter 2

Technical Background

> No one really knows what entropy really is, so in a debate you will always have the advantage.
>
> *John von Neumann, in discussion with Claude E. Shannon*

2.1 Introduction

Much of the work presented in this thesis is the result of the combination of two fundamental frameworks: that of *Markov decision processes*, and secondly that of *information theory*. The first is used to describe formally the process of an agent making decisions and acting in a possibly stochastic world. With this formulation, the second framework is used to model, analyse and optimise the informational properties of such processes, again in a formal way.

In this chapter I will give a brief introduction to these frameworks, and discuss the methods that have been developed in the respective fields. The aim of this introduction is to provide a sufficient and self-contained base on which to develop the methods of the coming chapters. Where needed, more specific background may be found in the following chapters. For a more in-depth review of MDPs and their solutions, the reader is directed to work by Sutton and Barto [97] and Bertsekas [18]; a thorough treatise of information theory is given by Cover and Thomas [31].

2.2 Information Theory

2.2.1 Basic Concepts and Notation

An important concept in probability theory, and by extension in information theory, is the *random variable*. This is a variable which can take on values from some set, called its *alphabet*, according to some probabilistic process. I will write random variables as capital letters, such as X, Y, W, and A, and their alphabets with the corresponding calligraphic capitals, $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{W}$, and $\mathscr{A}$. The size of an alphabet is called its *cardinality*, and is denoted as $|\mathscr{X}|$. Specific values of variables, also called *events*, will be written with the corresponding lower case letters, e.g. x, y, w, and a.

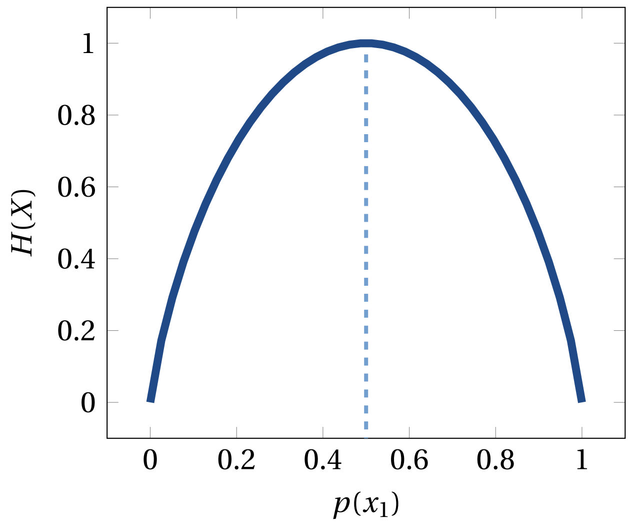


Figure 2.1: The entropy of a binary random variable X as a function of the probability of one of the two events, $p(x_1)$.

The probability distribution underlying a random variable is a function over all events in an alphabet, $\Pr(X = x)$, $\forall x \in \mathscr{X}$. To simplify notation, I will write $p(x)$ for the probability $\Pr(X = x)$ whenever it is clear which random variable is used. Whenever $p(x)$ is used without x specifically bound, I mean to refer (by abuse of notation) to the full distribution of X. In a similar vein I will write sums over the elements of alphabets, e.g. $\sum_{x \in \mathscr{X}}$, as $\sum_x$.

Some variables are 'more random' than others: a variable where it is as likely to see any of the possible events (picking a card from a well shuffled deck) shows more randomness than one that has one value with probability 1 (taking the top card of an ordered deck). This idea of randomness, or the *uncertainty* about the value of a random variable, is quantified by the information-theoretical concept of the *entropy* of a variable, $H(X)$, defined as follows:

$$H(X) = -\sum_x p(x) \log p(x) \tag{2.1}$$

The base of the logarithm determines the units in which entropy (and, as we will see later, information) is measured; I will use base 2, such that the units are *bits.*

As an example, take a binary random variable X with alphabet $\mathscr{X} = \{x_1, x_2\}$. In this case, the entropy is equal to $-p(x_1)\log p(x_1) - p(x_2)\log p(x_2)$. However, because probabilities add up to one, we know that $p(x_2) = 1 - p(x_1)$. Thus, we can plot the entropy as a function of $p(x_1)$, as is done in Fig. 2.1. As one would expect, this quantity is zero when $p(x_1) = 0$ and $p(x_1) = 1$, where there is no uncertainty about the outcome, and entropy is highest in the most uncertain case where $p(x_1) = p(x_2) = 0.5$.

One definition of information is as *the resolution of uncertainty.* In other words, knowing the value of a variable takes away the uncertainty about it. In that light, the entropy can then also be interpreted as the average amount of information you gain when you get to know a variable's value. If there are multiple variables, it could be that they are correlated, so that knowing the value of one reduces the uncertainty about the other. The average uncertainty still left about Y when you know the value of X is quantified by the *conditional entropy* $H(Y|X)$:

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) \leq H(Y) \tag{2.2}$$

The average reduction in uncertainty is then interpreted as the amount of information that knowing X gives about Y, which gives us the *mutual information* $I(X;Y)$ between the two variables:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \end{aligned} \tag{2.3}$$

The mutual information has some important well-known properties:

- It is non-negative: $I(X;Y) \geq 0$
- It is zero if and only if the variables are fully independent: $\forall x, y : p(x)p(y) = p(x,y) \leftrightarrow I(X;Y) = 0$
- It is symmetric: $I(X;Y) = I(Y;X)$
- It is upper bounded by the entropy of each separate variable: $I(X;Y) \leq H(X)$, $I(X;Y) \leq H(Y)$

Finally, if there is a third variable Z, one can determine how much information X gives about Y given knowledge of the value of Z, using the *conditional mutual information* $I(X;Y|Z)$:

$$\begin{aligned} I(X;Y|Z) &= H(Y|Z) - H(Y|X,Z) \\ &= \sum_z p(z) \sum_x p(x|z) \sum_y p(y|x,z) \log \frac{p(y|x,z)}{p(y|z)} \end{aligned} \tag{2.4}$$

The value of the this conditional mutual information depends on how the variables interact. Here an important concept is the *Markov chain*: three random variables X, Y, and Z are said to form a Markov chain $X \to Y \to Z$ if X and Z are independent given Y. This implies for the joint probability that $p(x,y,z) = p(x)p(y|x)p(z|y)$, and that $I(X;Z|Y) = 0$.

A central result in information theory is the so called *data processing inequality*, which states that in a Markov chain $X \to Y \to Z$, it always holds that $I(X;Z) \leq I(X;Y)$. The term comes from the interpretation of this inequality as showing that processing of data can only reduce information content: if the value of Y is obtained by processing X according to some function $y = f(x)$, no further processing of this value by a second function $z = g(y)$ can ever recover more data about the original input than what is retained in Y.

A corollary result is that in a Markov chain it always holds that $I(X;Y|Z) \leq I(X;Y)$. If the inequality is strict, it is said that there is *redundancy* between Y and Z: some of the information in X about Y is already given by knowing Z, so is redundant. This is certainly not always the case, however. If the variables do not form a Markov chain, it can be that $I(X;Y|Z) > I(X;Y)$. The canonical example of this is the XOR operation, where $Z = 1$ only

$$\begin{tikzcd} X \arrow[r] & Y \end{tikzcd}$$

Figure 2.2: Basic memory-less channel without feedback, with input X and output Y. The possible input and output values are described by $\mathscr{X}$ and $\mathscr{Y}$, and its behaviour by $p(y|x)$.

if the binary input variables X and Y differ in value. If the inputs are fully independent, knowing one gives no information about the other, so $I(X;Y) = 0$. However, if the output is given, the value of one is known exactly when knowing the other, and, assuming say that both X and Y are uniformly distributed, $I(X;Y|Z) = H(X) = H(Y) = 1$. This effect, where one variable 'unlocks' information between some other variables, is called *synergy*, and the intricate relationship between redundancy and synergy and their decomposition are currently the subject of active research and discussion [39].

A final important informational measure to introduce is the *Kullback-Leibler divergence* between two distributions, defined as:

$$D_{KL}\big(p(x) \parallel q(x)\big) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{2.5}$$

We will encounter this measure throughout this thesis. It gives some sense of distance between two distributions, being zero when they are equal, and growing if their probabilities diverge. However, it is not a proper metric, as it is not symmetric (generally, $D_{KL}\big(p \parallel q\big) \neq D_{KL}\big(q \parallel p\big)$). Also, it can be not well behaved: when $q(x) = 0$ for some x for which $p(x) > 0$ a division by 0 occurs.

Luckily these drawbacks do not appear in one important use case: when giving another definition of mutual information. One can namely show that:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D_{KL}\big(p(x,y) \parallel p(x)p(y)\big) \tag{2.6}$$

So, mutual information between two variables is the divergence between their joint distribution and the product of their marginals. This supports the second property of mutual information listed earlier: if X and Y are independent, i.e. $p(x)p(y) = p(x,y)$, the divergence, and therefor the mutual information, is 0.

2.2.2 Information Channels

'Static' information is not very interesting; we want information to 'move around'. When we use a telephone or the internet, information is constantly transferred from one point to another. Similarly, in an organism, information is passed around in its body: information about the environment is captured by its sensors, after which it can be transferred to its brain. Here it could be stored for later use, or it is passed on to its motor faculties to guide an action.

If information is passed from one place to the other, we say it travels through a *channel*. A channel is modelled with two random variables and their alphabets, that describe the possible inputs and outputs, and a transition probability function that defines its behaviour. For instance, the channel of Fig. 2.2 is defined by X, Y, $\mathscr{X}$, $\mathscr{Y}$, and $p(y|x)$.

Channels come in several flavors, determined by the possible interactions between the input and the output. In our basic example, the output of the channel only depends on the immediate input given to it; neither past inputs nor outputs have effect anymore: $p(y_t|x_t, y_{t-1}, x_{t-1}, \ldots, y_{t-k}, x_{t-k}) = p(y_t|x_t)$ for any $k > 0$. Such a channel is called *memory-less*, and if this equality does not hold the channel is said to *have memory*.

Alternatively, or additionally, it could be that a channel's output affects its next input. In this case we have that $p(x_{t+1}|y_t) \neq p(x_{t+1})$. For example, when calling over a bad phone line, you get information about the output on the other side when your friend says he cannot understand you. You can then select your input differently based on this information, perhaps by using simpler words, or by shouting. We say that such a channel has *feedback*.

2.2.3 Informational Variational Problems

With the concepts of the previous sections we are now ready to dive into what is the underlying process of most of the work in this thesis: the analysis of extremal properties of channels. I will look at interesting and necessary properties and structure of channels that operate at some minimal or maximal end of an informational spectrum. Specifically, I will try to find a channel's transition probability distribution that minimises and/or maximises some informational value(s), under some constraints.

Such problems fall under the mathematical field of calculus of variations: we optimise a mapping from probability distributions, or functions, to real numbered values, over the set of all permissible distributions. In this section I will summarise some of the existing work on such *informational variational problems*.

Rate-Distortion

One of the oldest informational variational problems is the *Rate-Distortion problem*[1]. It is described by Shannon in his seminal work [86]. All further concepts in this thesis can be seen as stemming from Rate-Distortion theory, so I will give a thorough introduction here.

The Rate-Distortion problem arises when you want to send data with as much compression as possible, while losing as little valuable information as possible. Imagine for instance when you want to transfer an image on a network, or onto a storage device, without compromising the quality too much. The compression can be arbitrarily large, up to where all information is lost, but it comes at a cost: losing more information makes the image ugly. Finally then, the question is: 'what is the least amount of information that needs to be preserved such that the cost does not get too high?'.

If X and Y are the input and output of a channel, and the cost of a mapping, called the *distortion* is D_p, then the rate-distortion problem is formulated as:

$$\min_{p(y|x)} I(X;Y), \text{ subj. to: } D_p \leq C_D, \tag{2.7}$$

[1] The other being the channel capacity problem.

where C_D is the maximum permissible cost/distortion[2]. A common way to define D_p is via a *distortion matrix* $\rho(x,y)$, that gives the cost of giving a certain output symbol for all input symbols. For instance, if you want a channel that simply reproduces all input, this function can be defined as:

$$\rho(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases} \tag{2.8}$$

In this case one wants to get the symbol right, and any wrong symbol is as bad as any other. A different example is when some symbols can be more wrong than others. Imagine that the symbols are integers and you think an output of 5 for an input of 2 is worse than an output of 1 or 3. In this case, you could alternatively use $\rho(x,y) = |x-y|$. To obtain the total cost, the expected value of ρ is taken:

$$D_p := E[\rho(X,Y)] = \sum_{x,y} p(x)p(y|x)\rho(x,y) \tag{2.9}$$

The problem of Eq. (2.7) is a constrained convex optimisation problem: the mutual information $I(X;Y)$ is convex in $p(y|x)$ for any specific x, and the constraint is a basic inequality constraint. The problem can be solved by using the method of Lagrange multipliers. First, the constrained problem is turned into an unconstrained problem by summing the function to minimise with the inequality constraint multiplied by a *Lagrange multiplier*, β. This gives the *Lagrange function* $\Lambda(p(y|x))$, and the following new problem:

$$\min_{p(y|x)} \Lambda(p(y|x),\beta) = \min_{p(y|x)} I(X;Y) + \beta E[\rho(X,Y)] \tag{2.10}$$

Given a fixed value of β the Lagrange function is still convex, for which the unique minimum can be found at
$\frac{\delta}{\delta p(y|x)}\Lambda(p(y|x),\beta) = 0$. Determining this derivative and solving for $p(y|x)$ finally gives the following self-consistent solution:

$$p(y|x) = \frac{1}{\mathcal{Z}} p(y)\exp[\beta\rho(x,y)], \tag{2.11}$$

where $\mathcal{Z} = \sum_{y'} p(y')\exp[\beta\rho(x,y')]$ is a normalising factor to ensure that $p(y|x)$ is a valid probability distribution and sums up to 1, and:

$$p(y) = \sum_x p(x)p(y|x) \tag{2.12}$$

As this solution is defined in terms of $p(y|x)$ itself, this does not give us the actual distribution yet. However, it turns out that if we set $\beta \in (0,\infty)$, start with any random guess of $p(y|x)$, and then iterate Eqs. (2.11) and (2.12), the distributions are guaranteed to converge to the solution of the original problem [31]. This method is known as the *Blahut-Arimoto algorithm*, after the two scientists who independently discovered it at the same time [20, 3].

[2] The minimisation is over all valid probability distributions, i.e. with the constraint that $\forall x, \sum_y p(y|x) = 1$. For readability, this constraint will not be mentioned explicitly any more in this thesis.

The value chosen for β determines the tightness of the distortion constraint: increasing it makes the constraint stronger, decreasing the amount of distortion. Conversely, decreasing β loosens the constraint, allowing to decrease the amount of information needed to be transferred, in exchange for an increase in distortion.

It is instructive to realise what happens at the limits of β. Firstly, the case where β approaches ∞ coincides with forcing the distortion to be at the absolute minimum. For our simple input copying channel example of before this would mean that $D_p := 0$. The solution gives a channel that achieves this constraint *while only transferring the least amount of information to do so.* At the other end, where β approaches 0, distortion becomes less and less important, allowing us to compress more and more, until finally all information is lost. Note however that even as $I(X;Y)$ approaches 0, this does not mean that the distortion can grow indefinitely, or that any information-destroying distribution is a valid variational solution. To see this, it may help to rewrite the problem as:

$$\min_{p(y|x)} I(X;Y) + \beta E[\rho(X,Y)] = \min_{p(y|x)} \beta' I(X;Y) + E[\rho(X,Y)], \tag{2.13}$$

where $\beta' = \frac{1}{\beta}$. So when $\beta \to 0$, $\beta' \to \infty$, and thus we can describe the solution as before: a channel that transfers the minimum amount of information (0 bits) *while only producing the least amount of distortion needed to do so.*

As an example, again consider our input copy problem, now with $\rho(x,y) = |x-y|$, and let $\mathcal{X} = \{0,1,2\}$ and $p(x)$ be uniformly distributed. It is intuitive to see that any mapping that gives different output distributions for different inputs transfers some information. On the other hand, if every input results in the same output, this output gives you no information at all about what went in. More generally, it holds that if each input gives the same *distribution* over outputs, the input and output are independent and no information is retained. To put it formally: for any solution where for all x, $p(y|x) = p(y)$, it holds that $I(X;Y) = 0$.

The resulting distortion for each result of this type however may not be the same. In our example, if all inputs are mapped to '0', i.e. $p(y=0) = 1$, we have for instance:

$$E[\rho(X,Y)] = \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 = 1. \tag{2.14}$$

Another solution that has $p(y)$ being uniform reduces this slightly, to $\frac{8}{9}$. However, the solution found using the algorithm described above with $\beta \to 0$ is $p(y=1) = 1$, which gives the least distortion, of $D_p = \frac{2}{3}$.

Traversing the full range of possible values for β gives a convex curve, as shown in Fig. 2.3 for our example. As said, at the extreme of $\beta \to 0$, we find zero information transmitted at the cost of a distortion of $\frac{2}{3}$. On the other end, when $\beta \to \infty$, we find that to have no distortion at all, about 1.58 bits of information must be transmitted. The exact amount is $\log 3 = H(X)$, in other words the full entropy of X.

Keep in mind that it is not always necessarily the case that all the information about the input needs to be transmitted to achieve minimal distortion. Above, we have enforced it through the distortion matrix, ρ. However, imagine we actually do not care about the output that is given for an input of '2', indicated by setting $\rho(2,y) = 0$ for all y. We can now map this input uniformly to either '0' or '1' without increasing distortion. The overall

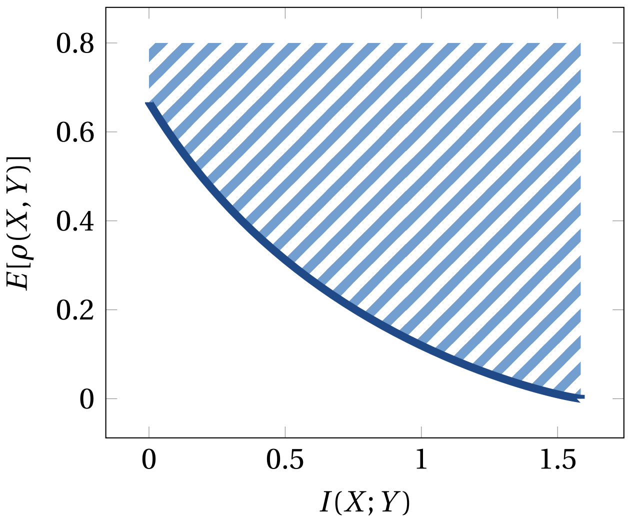


Figure 2.3: Rate-Distortion trade-off curve for the input copy problem. The convex curve traces the solutions for all possible optimal trade-offs, parameterised by $\beta \in (0,\infty)$. The banded area above the curve marks the feasible solutions: any channel distribution $p(y|x)$ results in a rate-distortion combination in this area. Or in other words: *no* solution can be found underneath the curve.

output is now 50% '0' and 50% '1', so $H(Y) = 1$. It turns out that in this case the rate that gives minimum distortion is even less than this, namely $I(X;Y) = \frac{2}{3}$. This is a result from the non-deterministic mapping, which prevents one to fully predict the output from the input, or to recover the correct input from the output; knowing one leaves some entropy about the other, which in this case results in $H(Y|X) = \frac{1}{3}$.

One last important thing to realise about the trade-off curve in Fig. 2.3, is that it separates all feasible solutions from the impossible ones. Any channel distribution $p(y|x)$ results in a rate-distortion combination on or above the curve; *no* channel can achieve a trade-off underneath the curve, and the curve is a *fundamental property* of the combined input distribution and distortion matrix.

Information Bottleneck

As we saw in the previous section, in rate-distortion theory, the relevance of information in the input is determined by the distortion matrix: information that does not influence the distortion is seen as irrelevant and discarded. This idea can be generalised to incorporate other indicators of relevance.

Imagine for instance that we have two random variables, X and Y, where knowing the value of X helps us to predict the value of Y. In other words, X holds information relevant to predicting Y, measured by $I(X;Y) > 0$. We can now construct a channel, with input X and output $\tilde{X}$, to transfer this, and only this information. So, similar to the rate-distortion problem, we aim to minimise the information transferred, $I(X;\tilde{X})$. However, instead of constraining distortion, we have the new condition that at least a certain amount of information C_I about Y must be retained. This means we have the following problem:

$$\min_{p(\tilde{x}|x)} I(X;\tilde{X}) \quad \text{subj. to} \quad I(\tilde{X},Y) \geq C_I \tag{2.15}$$

This is known as the *Information Bottleneck (IB)*, introduced by Tishby et al [104]. It derives its name from the observation that $\tilde{X}$ can be seen as a *bottleneck variable* through which information from X about Y is squeezed. A solution to the IB problem can be found in a similar way to the rate-distortion problem, by using a Lagrange multiplier to form a Lagrange equation and an unconstrained problem:

$$\min_{p(\tilde{x}|x)} \Lambda(p(\tilde{x}|x),\beta) = \tag{2.16}$$

$$\min_{p(\tilde{x}|x)} I(X;\tilde{X}) - \beta I(\tilde{X},Y), \tag{2.17}$$

and solve $\frac{\delta \Lambda(p(\tilde{x}|x),\beta)}{\delta p(\tilde{x}|x)} = 0$ for $p(\tilde{x}|x)$. This results in the following self-consistent solution:

$$p(\tilde{x}|x) = p(\tilde{x}) \exp\left[-\beta D_{KL}\left(p(y|x) \parallel p(y|\tilde{x})\right)\right] \tag{2.18}$$

With this solution, an iterative method similar to the Blahut-Arimoto algorithm can be performed to find solutions for the full range of $\beta \in (0,\infty)$. It is however important to note that although this algorithm will always converge, it may not do so to the global minimum of Eq. (2.17), since the original problem of (2.15) is non-convex. Several methods have been developed to find the optimum [91], in this thesis I will perform the iterative algorithm multiple times starting from different random initialisation of $p(\tilde{x}|x)$, to increase the probability of finding the desired solution.

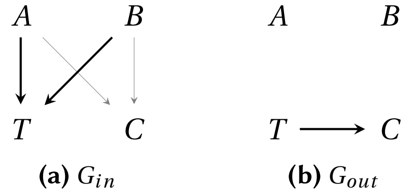
Multivariate Information Bottleneck

In the IB method as described above there is a single input variable, and a single relevance variable. This can be extended to scenarios where there are more variables with more complex interactions, where we can apply the so-called *Multivariate Information Bottleneck (MIB)* methods [92].

In this more general paradigm, the single variables are replaced by sets of random variables: given a set of random variables $\mathbf{X} = \{X_1, \dots, X_n\}$, we can construct a multivariate bottleneck with a set of bottleneck variables $\mathbf{T} = \{T_1, \dots, T_k\}$. These variables should capture the information in a selected input subset, $\mathbf{X}_{in} \subseteq \mathbf{X}$, that is relevant to a selected output subset $\mathbf{X}_{out} \subseteq \mathbf{X}$. This induces the new variational problem:

$$\min_{p(\mathbf{T}|\mathbf{X}_{in})} I(\mathbf{X}_{in};\mathbf{T}) \quad \text{subj. to} \quad I(\mathbf{T};\mathbf{X}_{out}) \geq C_I \tag{2.19}$$

One can represent this graphically to make it more intuitive. Figure 2.4 for instance gives a Bayesian network representation of one type of multivariate bottleneck. It consists of two networks: G_{in} which represents the original dependencies and which input variables the bottleneck variables should capture information from, and G_{out} which shows which variables this information should be relevant to. In this example, $\mathbf{X} = \{A, B, C\}$. Variables A and B predict C, so a reasonable choice for input variables could be $\mathbf{X}_{in} = \{A, B\}$, and $\mathbf{X}_{out} = \{C\}$. Here a single bottleneck-variable is used, so $\mathbf{T} = \{T\}$.



(a) G_{in} **(b)** G_{out}

Figure 2.4: Example of a multi-variate information bottleneck, where information from two variables, A and B, relevant to predicting C, is 'squeezed' through a bottleneck variable T. The given conditional dependencies are shown with thin grey arrows; those which are induced by the bottleneck, and are apparent in the final minimisation problem, with black arrows.

Let $\mathbf{U} = \mathbf{X} \cup \mathbf{T}$ be the set containing all random variables. Given the networks G_{in} and G_{out}, it can then be shown that the problem of (2.19) can be rewritten as:

$$\min_{p(\mathbf{T}|\mathbf{Pa}_{\mathbf{T}}^{G_{in}})} \sum_i I\left(U_i; \mathbf{Pa}_{U_i}^{G_{in}}\right) \quad \text{subj. to} \quad \sum_i I\left(U_i; \mathbf{Pa}_{U_i}^{G_{out}}\right) \geq \tilde{C}_I, \tag{2.20}$$

where the sums are over all variables in the networks, and $\mathbf{Pa}_X^G$ denotes the set of parents of variable X in graph G. This means that the information between variables connected in G_{in} is minimised, while maintaining a certain amount $\tilde{C}_I$ between those connected in G_{out}. The summations in (2.20) can often be further simplified, as usually several terms are constant with respect to $p(\mathbf{T}|\mathbf{Pa}_{\mathbf{T}}^{G_{in}})$. In the example of Fig. 2.4 this includes $I(X; A, B)$, symbolised by thin grey edges. This then gives the final problem as:

$$\min_{p(\mathbf{T}|\mathbf{Pa}_{\mathbf{T}}^{G_{in}})} I(T; A, B) \text{ , subj. to: } I(C; T) \geq \tilde{C}_I, \tag{2.21}$$

To solve such a problem, we can again take the familiar approach, of creating an unconstrained problem by introducing a Lagrange multiplier, set the derivative of the resulting Lagrange equation to zero, and solve for $p(\mathbf{T}|\mathbf{Pa}_{\mathbf{T}}^{G_{in}})$. For our example this results in:

$$p(t|a,b) = \frac{1}{\mathcal{Z}} p(t) \exp\left[-\beta D_{KL}\big(p(c|a,b) \parallel p(c|t)\big)\right] \tag{2.22}$$

2.3 Markov Decision Processes

The framework of Markov decision processes (MDPs) is used to formulate an agent-environment system and its interactions. The formulation of an MDP describes how the state of a system develops over time and is transformed as a result of decisions made by an acting agent. It also describes the costs of *performing* the actions as an outcome of the available decisions, and as such opens up the problem of finding a decision making policy that minimises the amount of this cost incurred over time.

The following subsection will give the formal description of an MDP, and introduces more of the notation that will be used in the remainder of this thesis. The rest of this section will introduce some standard methods for solving MDPs.

2.3.1 Formulation

An MDP can be broadly described as a formal description of a sequence generated by a non-linear recurrence relation [17]. In the context of agent-environment systems, this sequence consists of a series of costs incurred, or rewards received, by a decision making agent.

Let the set of possible states that the agent-environment system can be in be $\mathscr{W}$ (the *state set*). In the remainder, it is assumed that the state set is finite, and that both time and state are discrete. At a given time t, the system is in a specific state $w_t \in \mathscr{W}$, and the agent chooses and performs an action a_t, from its set of available actions $\mathscr{A}$ (the *action set*). The evolution of the system's state, given the current state and the chosen action, is determined by the *state transition probabilities*:

$$P^a_{ww'} = Pr(w_{t+1} = w' | w_t = w, a_t = a) \tag{2.23}$$

In this framework, it is assumed that the agent receives direct feedback about the actions that it takes. This feedback comes in the form of some quantifiable *reward*, or conversely some *cost* which is modelled with negative reward, which it receives in the time step after performing an action. The amount of expected direct reward r_{t+1} received at some time $t+1$ (after acting in time t), is determined by a reward function:

$$R^a_{ww'} = E[r_{t+1} | w_t = w, a_t = a, w_{t+1} = w'] \tag{2.24}$$

Given these concepts, an MDP is then defined by a tuple $< \mathscr{W}, \mathscr{A}, P^a_{ww'}, R^a_{ww'} >$.

The problem for an agent that makes decisions and acts inside such a process, is to do this in such a way as to maximise the expected reward received, or minimise the expected cost incurred by it. Its decision making process is modelled by a *policy*, π, which gives the probability of performing an action given a certain state of the world:

$$\pi(a|w) = Pr(a_t = a | w_t = w) \tag{2.25}$$

Note that the policy in this thesis is assumed to be time-invariant: after setting on a policy, the probability of selecting an action when in a certain state is the same for any time t. In parts later chapters I will treat learning agents, of which the policy does change over time, however in all methods this is ignored and time-invariance is still assumed. As will be shown, this does not prevent those methods from arriving at important results.

Together with the MDP definition, the policy fully defines the expected sequence of generated reward. Let $V^\pi(w)$, the *value function*, be the function that gives the expected sum of the rewards in such a sequence, starting with the system in state w and following policy π, then it holds that [97]:

$$\begin{aligned}
V^{\pi}(w) &= E\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|w_t = w,\pi\right] \\
&= E\left[r_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|w_t = w,\pi\right] \\
&= \sum_a \pi(a|w)\sum_{w'} P^a_{ww'}\left[R^a_{ww'} + \gamma E\Big[\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|w_{t+2} = w',\pi\Big]\right] \\
&= \sum_a \pi(a|w)\sum_{w'} P^a_{ww'}\left[R^a_{ww'} + \gamma V^{\pi}(w')\right] \qquad (2.26)
\end{aligned}$$

The last, recursive form is commonly called a *Bellman equation*, after one of the first authors to describe MDPs, Richard Bellman [17]. The factor $\gamma \in [0-1]$ is a *discount factor.* It indicates the importance of future expected reward compared to short term reward: if γ is large, future reward is weighed high in the expected sum, but if it is low, immediate rewards have a much higher relative weight. In a pragmatic sense, it is inserted to prevent the expected sum from becoming infinite. Formally, one can treat it as the probability of the agent 'surviving' and seeing the next time step, an interpretation that will emerge in chapter 5.

2.3.2 Solving MDPs

Since their first formulations in the 1960s, a large amount of work has been performed into solving MDPs. Finding a solution comes down to finding a policy that maximises the expected future sum of reward, which is equivalent to solving:

$$\pi^* = \arg\max_{\pi} V^{\pi}(w), \qquad (2.27)$$

where π^* is called an *optimal policy.*

Such a solution is not necessarily unique. The number of solutions can even be infinite: if there are two optimal policies π^*_1 and π^*_2, any convex combination, $\alpha\pi^*_1(\cdot|w) + (1-\alpha)\pi^*_2(\cdot|w)$ with $\alpha \in [0,1]$, results in a new optimal policy. When we do have a set of policies that result in the same expected reward, the question arises of whether one is preferred over the other. This thesis centrally discusses one possible source of preference: the cognitive requirements of a policy, or the cost of *making* decisions. Before we delve into that, however, we look at some traditional methods for solving MDPs, which inspire methods developed and used later on.

Value Iteration

Instead of finding an optimal policy directly, one can first find the *optimal value function*:

$$V^*(w) = \max_a P^a_{ww'}\left[R^a_{ww'} + \gamma V^*(w')\right]. \qquad (2.28)$$

If the full MDP is known, this optimal value function solution can be found through a process called *value iteration*, which is formed by simply iterating Eq. (2.28), until the value function converges. It can be shown that this process is guaranteed to converge to the globally optimal value function.

If we then define the *state-action value function*, which from here will be referred to as the *utility function*, as:

$$U^{\pi}(w,a) = \sum_{w'} P^{a}_{ww'} \left[R^{a}_{ww'} + \gamma V^{\pi}(w')\right] \tag{2.29}$$

$$= \sum_{w'} P^{a}_{ww'} \left[R^{a}_{ww'} + \gamma \sum_{a'} \pi(a'|w') U^{\pi}(w',a')\right], \tag{2.30}$$

with the optimum $U^*(w,a) = \max_\pi U^\pi(w,a)$, one optimal policy, which maximizes action entropy (see the next chapter for discussion of such a choice) is then found by choosing:

$$\pi^*(a|w) = \begin{cases} \frac{1}{n(w)} & \text{if } U^*(w,a) = \max_{a'} U^*(w,a') \\ 0 & \text{otherwise} \end{cases}, \tag{2.31}$$

where $n(w) = \left|\{a : U^*(w,a) = \max_{a'} U^*(w,a')\}\right|$, i.e. the number of optimal actions for state w.

Reinforcement Learning

Value iteration is an *off-line* algorithm. This means that an agent can perform it before it starts to find an optimal policy, and use this outcome to *plan* its actions ahead of time. In contrast, an *on-line* algorithm is used when an agent learns the value of states (and actions) as it is interacting with the environment. Such an approach is necessary if the agent does not know either or both of the reward and transition function.

A popular example of such an on-line algorithm is *Q-learning*. Using this method, the agent maintains an estimate $\tilde{U}(w,a)$ of the utility function, which gives the utility that an agent predicts. At any time $t+1$, however, the agent can also make an empirical estimate based on a combination of the actual observed reward and the maximum estimated future utility, $r_{t+1} + \max_{a_{t+1}} \tilde{U}(w_{t+1}, a_{t+1})$. The learning then consists of updating the utility estimate to be closer to the observation, following the rule:

$$\delta = r_{t+1} + \gamma \max_{a_{t+1}} \tilde{U}(w_{t+1}, a_{t+1}) - \tilde{U}(w_t, a_t) \tag{2.32}$$

$$\tilde{U}(w_t, a_t) \leftarrow \tilde{U}(w_t, a_t) + \alpha\delta. \tag{2.33}$$

The difference δ is known as the *prediction error*, and the parameter $\alpha > 0$ is the *learning rate.* A higher rate can cause faster learning, but may cause large oscillations of the learned values, which may hinder convergence. A common strategy is to start out learning with a high learning rate, and slowly decrease it over time. This learning method will converge on the optimal solution if the rate is decreased slow enough [97].

Besides being on-line, Q-learning also is an *off-policy* method. This means that it will converge to an optimal solution, without requiring that the agent follows the policy that it deems optimal at all times. This helps an agent prevent getting stuck in a local optimum: instead of sticking to what it knows, given by its value estimate, it can try out random actions to see if there is a better solution, without polluting its hitherto gained experience.

In theory it could even perform a completely random walk, but in practice it is more efficient to limit the randomness. One approach is to follow a policy which is optimal with respect to the current value estimate, but with probability $\epsilon \in (0, 1]$ choose a random action instead. Such a policy is called $\epsilon - greedy$. Another possibility is to use a *soft-max approach*, where the policy is defined as:

$$\pi(a|w) = \frac{1}{\mathcal{Z}} \exp[\beta \tilde{U}(w, a)], \tag{2.34}$$

where $\mathcal{Z} = \sum_{a'} \beta \tilde{U}(w, a')$ is a normalisation factor. The parameter β functions as an inverse temperature: when set to 0, behaviour is fully random, and action selection becomes more greedy with growing β. In the limit $\beta \to \infty$, this equation becomes equivalent to Eq. (2.31). As we will see later, this type of exponential form is arrived at naturally when combining Information Theory with MDPs.

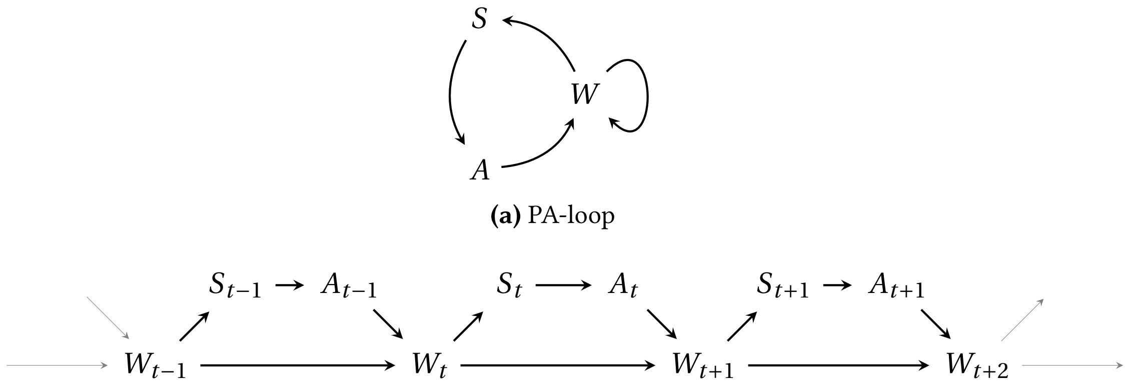
2.4 Partially Observable MDPs

The MDP formulation makes the assumption that the full state of the world is available to the decision making agent. In other words, the world is *fully observable*. This assumption does not always hold true, especially not if we consider agents operating in a real environment, embodied with limited sensors, such as animals, physical robots, or ourselves. We can only sense our immediate surroundings, and cannot observe directly the state of things behind walls, or the mental state of other agents. In such a case, the world is said to be *partially observable*.

In the MDP framework, this can be modelled by having a policy $\pi(a|s)$ that depends not on the world state w directly, but on a possibly incomplete observation s of that state by the agent's sensors. The possible observations are given by the sensation set $\mathscr{S}$, and the sensing process is described by a probabilistic sensing model $p(s|w)$. An MDP with this extension is referred to as a *partially observable MDP*, or *POMDP*. For a review of solutions to POMDPs, and usage examples, the reader is directed to work by Kaelbling et al [43] and Littman [60].

2.5 Informational Treatment of MDPs

The interactions between an agent and the environment form a closed loop, the *perception-action loop (PA-loop)*. The world is in a state, part of which is captured by the agent's sensors. Based on this, the agent selects an action, which is combined with the inherent development of the world to determine the next world state, which closes the loop. This loop is graphically sketched in Fig 2.5a. In the POMDP description, these interactions are stochastically and fully described by the transition probabilities $P^a_{ww'}$, $\pi(a|s)$, and $p(s|w)$. This means that the states of the system can be represented as random variables: the world state at time t is represented by W_t, and the agent's sensor state and its selected action by S_t and A_t, respectively. By unrolling the PA-loop in time and constructing a graph with the random variables as vertices and their interactions indicated with edges, one arrives at a *causal Bayesian network (CBN)* description of the system: a directed acyclic graph



(a) PA-loop

(b) CBN

Figure 2.5: The perception-action loop (PA-loop). (a) The interactions between an agent and the environment can be graphically represented with a loop: the world is in a state W, part of which is captured by the agent's sensor S. Based on this, the agent selects an action A, which is combined with the inherent development of the world to determine the next world state, which closes the loop. (b) When the PA-loop is unrolled in time, and its elements are represented by random variables, a causal Bayesian network as a formal description of the system is constructed.

where the edges coming into a vertex are from its immediate causes, and the joint probability over all variables is the product of the distributions of all variables conditioned on their direct parents [73]. The CBN of a POMDP is shown in Fig. 2.5b. In a fully observable scenario, the world and sensor states collapse, and the sensor variables can be replaced with a direct link from W_t to A_t.

The edges entering the random variables in a CBN can be thought of as information channels: information about the world is transferred through the sensory channel, which is processed in the agent's cognitive channel resulting in an action selection, and this action induces an actuation channel that governs the transition of the world state [107, 51]. In this view, one can apply the information theoretical concepts described previously to MDPs, opening up a large collection of tools, formalisms and proofs to the study of decision making.

In the following section I will discuss the fundamental concept for this thesis which arises from applying information theory to behaviour: *relevant information.*

2.6 Relevant Information

In the informational view of the PA-loop, one can formally quantify the amount of world information that is on average gathered by the agent through its sensor with the mutual information $I(W;S)$. The mutual information $I(S;A)$ gives how much of this sensed information is on average actively used in the action selection process. As mentioned before, in the fully observable case, which I will adhere to for now and for most of first part of this thesis, the sensor and world state collapse, and one can regard $I(W;A)$ directly as the amount of information that an agent on average takes in and processes per time step.

This amount depends on the agent's policy: the mutual information is convex in $\pi(a_t|w_t)$. In other words, different policies require different informational bandwidths, or different information processing capacity in the agent. It is argued that this required capacity can be correlated to the metabolic cost of information acquisition and processing, and as such constitutes a quantitative measure of cognitive burden [76]; an argument that is fundamental to the work in this thesis.

The parsimony pressure put forward in the previous chapter, then would act to minimise this quantity, while the efficiency pressure drives towards higher performance. With the formal framework of the previous sections, one can now make this trade-off precise. To determine the minimum bandwidth that is required to achieve at least a certain level of performance, as measured by the expected utility described above, one seeks to solve the following minimisation problem:

$$\min_{\pi} I(W;A) \quad \text{subj. to} \quad E[U^{\pi}(W,A)|\pi] \geq C_U \tag{2.35}$$

The mutual information found as the solution of this problem is dubbed the *relevant information (RI)* [77]: similar to how an information bottleneck aims to only capture the information that it is relevant to predicting the output variable, solving (2.35) gives the amount of information that is strictly relevant to successfully achieving the required performance. Any other amount can be discarded, to alleviate the costs of requiring a larger bandwidth.

To solve the problem, a similar approach is taken as with the earlier variational problems. Firstly, a Lagrange equation is constructed:

$$\Lambda(\pi,\beta) = I(W;A) - \beta E[U^{\pi}(W,A)|\pi] \tag{2.36}$$

Next, one finds the partial derivative of this equation:

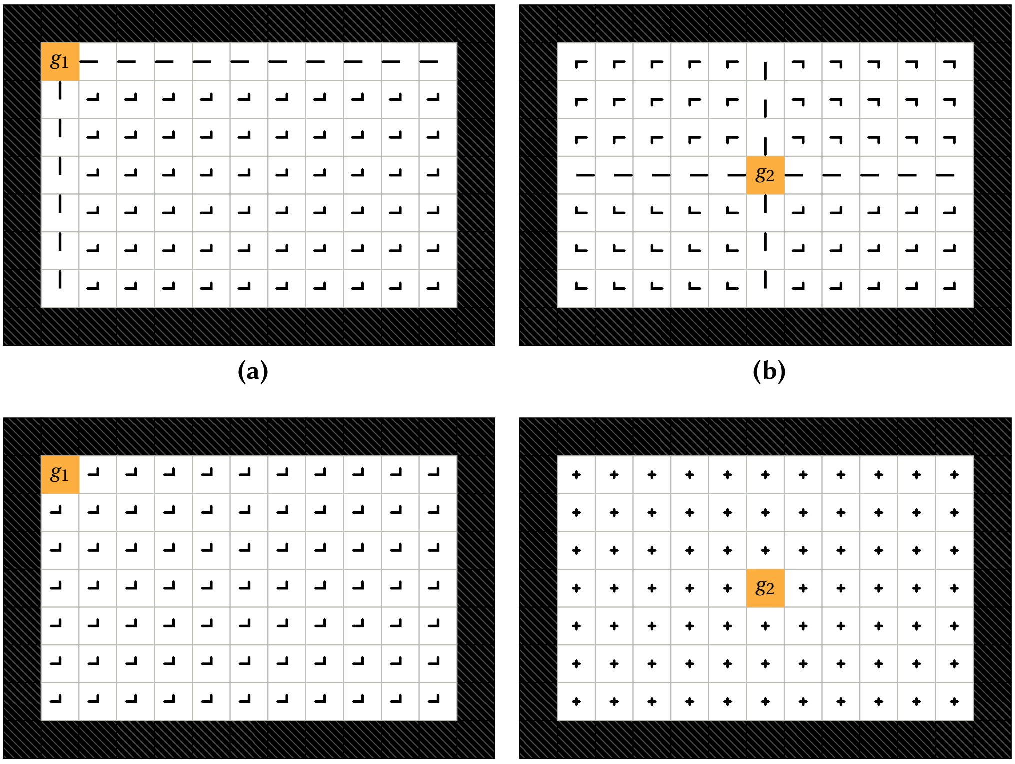
$$\frac{\delta}{\delta\pi}\Lambda(\pi,\beta) = p(w)\log\frac{\pi(a|w)}{p(a)} - \beta p(w)U(w,A) \tag{2.37}$$

Finally, this derivative is equated to zero, which after some rearranging of terms gives the self-consistent solution:

$$\pi(a|w) = \frac{1}{\mathcal{Z}}p(a)\exp[\beta U^{\pi}(w,a)] \tag{2.38}$$

This solution has a form very similar to that of the rate-distortion problem, however with one important distinction: the distortion measure, i.e. the cost, or negative utility, here *does* depend on the mapping, given by the policy, that the minimisation is done over. When an iterative algorithm is constructed in order to find the final solution, using (2.38) to update the policy in each iteration, one must ensure that the utility is consistent with the policy.

In the relevant information method this is done by interleaving the updates according to (2.38) with value-iteration updates of the utility function according to (2.30). Performing these iterations until convergence of both the policy and the utility function results in the final algorithm listed as Alg. 1. Currently no convergence proof of this algorithm consists, but in practice convergence is good.



(a) **(b)** **(c)** **(d)**

Figure 2.6: Example grid-world navigation problem. At any time the agent finds itself in one of the open cells, which determines the fully observable state. The task for the agent is to choose actions based on its observed location in such a way as to navigate towards a goal state, where it receives a reward of 1. At all other steps the reward is 0. In the left two figures the goal, g_1 is in the north-west corner, in the right two the goal, g_2 is in the centre. The top two figures show the policies resulting from applying the RI method with $\beta \to \infty$, whereas the bottom two show those acquired with $\beta \to 0$, where the length of the lines from the center of the cells is proportionate to the probability of selecting the action that would move the agent into that direction.

Algorithm 1 Relevant Information

```
Require: $<\mathscr{W},\mathscr{A},P^a_{ww'},R^a_{ww'}>,\beta,\gamma$
$p(w) \leftarrow \frac{1}{|\mathscr{W}|}$
$\pi_0(a|w) \leftarrow \frac{1}{|\mathscr{A}|}$
$U_0(w,a) \leftarrow 0$
$k \leftarrow 0$
repeat
    $p_k(a) \leftarrow \sum_w p(w)\pi(a|w)$
    $\pi_{k+1}(a|w) \leftarrow \frac{1}{Z} p_k(a) \exp\left[\beta U_k(w,a)\right]$
    $U_{k+1}(w,a) \leftarrow \sum_{w'} P^a_{ww'}\left[R^a_{ww'} + \gamma \sum_{a'} \pi_k(a'|w') U_k(w',a')\right]$
    $k \leftarrow k+1$
until $D_{JS}(\pi_k \parallel \pi_{k-1}) \leq \epsilon_\pi$, $\parallel U_k - U_{k-1} \parallel_\infty \leq \epsilon_U$
```

As an example to show the properties of relevant information, consider a navigation task in a simple grid world, shown in Fig. 2.6. At any time, the agent is located in one of the 7 × 14 cells, and this location gives the state of the world W_t, which is fully observable to the agent. Based on these observations, the agent can choose one of four actions: move north, east, south or west. Transitions are noiseless, so the agent will always move to the cell targeted by its action, except for when it hits the surrounding wall. In this case the agent stays in the same cell. The agent's task is to reach a goal cell, where it receives a reward of 1. At any other step the reward is 0. In Figs. 2.6 two possible goals are shown: one in the north-west corner (g_1), and one in the centre of the world (g_2). For the remainder, $\gamma = 0.95$ is chosen.

One can now apply the relevant information method to these two scenarios. Firstly, we want to set $\beta \rightarrow \infty$, to acquire the policies that achieve maximum utility with the lowest bandwidth. Actually moving towards infinity is not feasible, in practice a high enough value is chosen such that increasing it further does not change the outcome more than some small value ϵ. Choosing a value *too* high however can make the computed exponents diverge and cause numerical instability. For this example, a choice of $\beta = 10^6$ is a good trade-off, and is chosen as well as for experiments in the remainder of this thesis, unless noted otherwise.

The policies obtained for this value of β are shown for both cases in Figs. 2.6a and 2.6b, with bars whose lengths are relative to the probability of choosing the corresponding action. Some properties of the policies are visible, that are commonly seen in RI solutions:

- In states where the same set of actions achieves the required utility level, the distribution over these actions is the same. In contrast, this is not a necessary result in traditional MDP solutions, since different local policies do not change the final expected reward. When taking into account the informational cost however, the result seen here is desired: due to this, all these states can be treated together, and the agent does not need to make a distinction between them. Making distinctions requires information, and if the local policy at some state deviates, additional information is required to check if this state is the current one.

- The common distribution of the previous point is a particular, skewed one. The

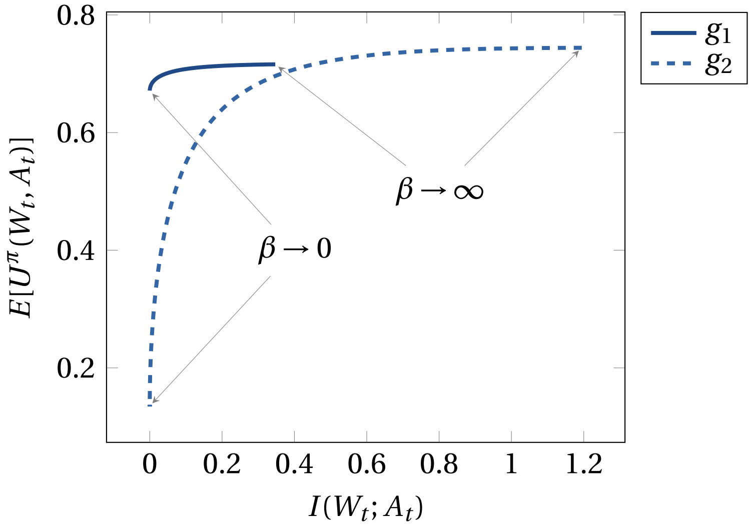


Figure 2.7: Trade-off curves obtained with the RI method for the scenarios of Fig. 2.6.

probability of moving sideways is $\frac{5}{8}$, against a probability of $\frac{3}{8}$ of moving along the north/south axis. It is no coincidence that this distribution matches the layout of the world. Specifically, this distribution is shaped by the states where the agent has no choice: those in a straight line with the goal state. The policy in the other states is chosen to match the distribution over the actions in these critical states as closely as possible.

Both properties can be summarised as one drive in the relevant information method: make the local policy $\pi(a_t|w_t)$ as close to the average policy $p(a_t)$ as possible. Informationally, this causes the conditional entropy $H(A_t|W_t)$ to be as close to the a-priori entropy $H(A_t)$ as possible, the difference of which exactly is the mutual information $I(W_t; A_t)$. The value of this difference, and thus the relevant information, differs significantly for the two scenarios: 0.35 bit in the case of g_1, and 1.19 bit for g_2. Moving into a corner is informationally much easier than exactly hitting a cell in the centre; there are only 2 actions to choose between, and in a large part of the world the choice probabilities are the same, whereas in the other scenario the agent has to discern with much more detail where it is relative to the goal to select the correct action.

Next, we will compare the other end of the trade-off curve, where $\beta \to 0$, and see that this difference in difficulty can also be seen here. As before, moving indefinitely towards the limit is not feasible, so in practice β is reduced until convergence of the found trade-off. Note that setting $\beta = 0$ would fully cancel out the utility condition, allowing any zero-bandwidth policy, which is not what we are after: we want the one that gets the best performance with (, in practice, very close to) no information. Figures 2.6c and 2.6d show the policies obtained for both scenarios in this case. For both, the relevant information is 0, but the expected utility differs greatly: 0.67 for g_1, against 0.13 for g_2. The difference in policies is clear: in the first scenario the policy can still be directed, such that even a blind agent can still hit the goal by randomly going north or west, whereas in the second scenario the agent has no such guidance, and can do no better than a random walk.

The full trade-off curve is traced by moving the value of β along the range $(0,\infty)$. These curves for the two scenarios above are plotted in Fig. 2.7. Again, the significant informational difference is apparent, a difference that would not be apparent in traditional machine learning treatments that focus solely on reward. These examples hint towards recurring principle in this thesis: *Exploitable structure in the perception-action loop supplies an opportunity to alleviate cognitive burden.*

There is one aspect of the RI method I have not discussed yet: the choice of the state distribution $p(w_t)$. This distribution strongly influences, together with the policy, the values of $p(a_t)$, $I(W_t;A_t)$, and $E[U^{\pi}(W_t,A_t)]$. Here, inline with the original formulation of the RI method [77], the uniform distribution is chosen. This corresponds to the assumption that the agent has no prior knowledge about which state is more likely to be observed, and where it has no knowledge about how its actions determine this distribution, or how a change in behaviour could shift it. In chapter 5 I will investigate the result of dropping this assumption.

Chapter 3

Related Work

3.1 Information and Cognition

The work in this thesis is based on the viewpoint of an agent as an information processing system, and that the fundamental aspects of such a system can be studied using the tools of information theory. This approach was already adopted by some early pioneers shortly after the birth of the field, and has gained in popularity in the last decades. In this section, I will provide an overview of work following this approach, and discuss the relation to the work presented in this thesis. I will start with its use in the study of biological cognition, which has a slightly longer history, followed by its application in computational domains, noting however that there often is some overlap between the two.

3.1.1 Biology and Neuro-Science (wet information)

Some of the earliest applications of information theory to cognition I have already mentioned earlier in the introduction. The new field created by Shannon was recognised by Attneave [6], Watanabe [111], and Barlow [12] to provide a concrete grasp on an idea that had been put forward in more abstract form already a century earlier [11]: that the statistics of sensory input, and in particular the statistical redundancies in this input, are important for cognition. The main hypothesis of Attneave and Barlow was that of informational economy: it is expected that the perception and cognition of an organism is no more expensive than it needs to be, and that this expense can be measured using the tools of information theory, which is very close to the Parsimony hypothesis.

This hypothesis was picked up in the work of a range of authors, which was reviewed by Attick [5]. In his review, he presented the principle trade-off of coding to improve efficiency without losing information as a variational minimisation problem, which can be seen as an early form of the information bottleneck method introduced by Tishby et al. [104], and is very similar to ones that I will use to study the trade-off between parsimony and performance.

As discussed before, one prediction stemming from the economy hypothesis was that informational redundancy is minimal in sensors and the brain. It now turns out that this prediction does not accurately describe reality, but as Barlow noted, this does not invalidate the original economy hypothesis, and the informational study of redundancy is still a worthwhile endeavour [11]. In any case, these results have shown that the application of information theory to perception and cognition can supply quantitative predictions, which can be measured and tested in biology, bringing a wide range of formal tools and laws into the field.

The strength of this approach is recognised by Borst and Theunissen [21], who state that information-theoretical comparisons 'validate assumptions present in any neurophysiological analysis'. They go on to present an overview of the use of methods from information theory to make quantitative observations of neural processes, showing how to measure information transfer in practice directly, or through lower or upper bounds. Such results provide empirical grounding for the more theoretical work in artificial, computational systems; work as reviewed in the following section, and as will be performed in the remainder of this thesis.

Although in general the use of information theory as an observational tool in neuroscience seems accepted, the question of which informational measurement is best to use may still spark some discussion. One fundamental question is how to treat individual events. Informational quantities such as entropy and mutual information are averages over the full distribution of possible events. Decompositions of these quantities on a per-symbol basis may not be well-behaved. For instance, the entropy $H(X)$ is the weighted average of the symbol specific 'self-information' $-\log p(x)$, which is undefined for $p(x) = 0$. Several measurements have been proposed for information in a single impulse or response, based on different properties that such a measure is required to have [35, 24].

In Sec. 5.2.2 I will discuss these, and how to choose between them in more detail. One shared property however is that they derive from the mutual information between impulse and response, as most of the used measurements are (c.q. again Borst and Theunissen [21]). Sinanović and Johnson however put forward that to measure how well systems actually *process* information, mutual information is not a good choice [90]. Their argument is based on the claim that it is inherently impossible to determine the distribution from which 'information', is drawn. If one overlooks their seeming confusion of raw symbols with the information that is carried by them, this comes down to saying that it is not possible for an experimenter to present stimuli to a system, distributed exactly as how the system would receive them 'in the wild'. Since the mutual information relies heavily on the input distribution, measurements of this quantity can be influenced heavily by the choice of inputs by the experimenters, and may not be related to actual information processing. To overcome this, the authors develop a measurement called the information transfer ratio, which aims to quantify, using the Kullback-Leibler divergence, the response to *changes* in inputs.

In the artificial scenarios studied in this thesis, I am not affected by the practical difficulties encountered by a neuro-scientific experimenter, that motivated these authors. I argue that for the study performed here, the mutual information between input and output is a valid measure for what the information transfer ratio was designed for: to what precision are changes in input reflected in the output? Furthermore, in the experiments presented here, and in those of other authors as reviewed in the next section, one *is* able to determine the full joint distribution of input and output.

This being said, the observation of the importance of the input distribution is instructive to keep in mind, and underpins the work presented in chapter 5. Furthermore, two other foundations mentioned by the authors also underlie the work in this thesis:

1. "What may or may not be information is determined by the information sink." This

way of phrasing may again confuse data with information, in the vocabulary of built up in the previous chapter this translates to: 'Information is measured by Shannon information, what of this information is *relevant* is determined by performance.'

2. "When systems act on their input signal(s) and produce output signal(s), they indirectly perform information processing [and] the result of processing information is an action." This is in line with the 'agent as information processing system' view adopted in this thesis.

3.1.2 Artificial Agents (dry information)

Most work in the previous section focuses on modelling and studying properties of biological systems, using information-theoretical concepts mainly to perform measurements. Some of it already showed one opportunity given by applying information theory: it can give quantitative predictions about cognition. More strongly, its formal laws and methods allow one to formulate and study hard bounds on perception, cognition, and action. This is the main motivation of grounding the work of this thesis in information theory, as it has been done for a body of work before, which I will review here. The main approach is as follows: formulate an informational concept that is of importance to a system, and apply the well established methods of information theory to study it.

As an example, the study of information-theoretical limits of control by Touchette and Lloyd [108, 107], and their models of a control loop, strongly underpin the work in this thesis. They formulated controllability, observability, and stability of a control system using purely information-theoretical concepts, which allowed them to derive formal bounds on these concepts. Tatikonda and Mitter performed similar studies on systems with feedback [100].

I will employ similar models, however placed in the slightly different domain of acting agents. A main theme in this type of work, including the work presented here, is how information is organised in an agent, and how properties of information can help self-organisation. Topics related to this theme include how informational structure is imposed on sensory input by embodiment [74], and how this information is further shaped by coordinated and dynamic interactions with the world [61]. Such work supplies a formal foundation to the idea that embodiment is a crucial aspect of cognition, and show that studying the informational interaction between agents and their environment can provide important insights.

For instance, one can consider an acting agent as 'injecting' information into the environment, and the more it can inject, the more control it has on the environment. The informational concept of *empowerment* [52] determines the upper bound on the *observable* control that an agent can exert, in contrast to control that is unobservable and therefor irrelevant to the agent. Empowerment is formulated as the maximum capacity of the information channel from an agent's actions to its future sensory states. Posing maximisation of this quantity as an intrinsic drive results in interesting self-organisation, such as internal representations of space and time [49], emergent pole balancing [50], forming an *Umwelt*[26], and 'meaningful' pathways through an environment [2]. The maximisation of possible actuation bandwidth as done to maximise empowerment is in some sense dual to the approach taken in this thesis, where I *minimise* the *required sensory* bandwidth, which similarly, but differently, gives rise to self-organised aspect of cognition and behaviour.

Other examples of self-organisation driven by informational optimisation include infotaxis [110], where a drive to acquire information about a target can be sufficient for an agent to achieve it in a biologically plausible way, and organisation of behaviour to maximise the information between the past and the future [8]. This latter quantity, *predictive information*, has been posed to equate the complexity of a system, and to indicate bounds on the learnability of a model of the system [19].

Information theory has been applied to other facets of learning as well. It provides a minimum bound on the information that is required to learn a task simultaneously with other tasks [14], or it is used to develop novel learning algorithms, such as by applying the information-theoretical methods of an information bottleneck [85]. Introducing information-theoretical costs on actions, to punish 'large' actions that deviate a lot from doing nothing, can even transform learning problems such that they become linear and can be solved analytically, rather then through exhaustive search [106], though under the assumption that the agent can freely manipulate the complete dynamics of the agent-environment system.

Exploration is another aspect of learning for which information theory readily supplies tools to study it. The simplest way to perform exploration is through sheer random action selection. A popular concept in this is that of *maximum entropy* [41]: given knowledge gathered so far of a system, from the best fitting models choose the one with the highest entropy. This is the model with the least commitment, and in a learning acting agent leads to maximum randomisation of actions and thus exploration. With exploration equated to action entropy, one can even trade off exploration with performance in a principled way [81], giving rise to methods similar to those I will employ to trade off performance with sensory bandwidth. Another learning application of maximum entropy is in reverse reinforcement learning, where an agent learns to solve a task by observing how another (human) agent performs solutions [114].

Other approaches that arrive at similar methods that also induce intrinsic exploration are based on maximising predictive information [7], and maximising information about the dynamics of the environment through interaction with it [94], which provides an information theoretical model of curiosity that may drive reinforcement learning [95].

Much of these approaches involve trade-offs between different drives and costs, which also forms the foundation of this thesis. Another closely related example of such a trade-off is that between a learner's performance, and the capacity needed to store solutions [102]. Whereas this work shows how to derive necessary bounds on the static information *storage* capabilities of an agent's cognition, I will perform similar studies on the *dynamic* information *processing* requirements, starting with a similar hypothesis and arriving at an analogous outcome: an agent prefers informational minimisation, which gives rise to inherent structuring.

The work in this thesis will only involve single agents, however the approach of information optimisation can also be applied to multi-agent scenarios. Maximising structure in the available information for instance aids the evolution of coordinated behaviour [93], whereas predictive information maximisation can lead to cooperation [34]. Interestingly, coordinated behaviour can also arise when agents perform purely egocentric information *minimisation*. The concept of *digested information* [83] tells us that due to the symmetric properties of information, an agent that minimises the information that it takes in to the relevant minimum, also *injects* this information back into the world in a highly densified form through its actions. This motivates other relevant information seeking agents to stay within range of this agent, and induces coordination of their actions, giving rise to a form of flocking grounded in egocentric drives.

3.2 Further Related Work

The work in this thesis is partly related to behavioural ecology: the study of the consequence of behaviour on fitness [54]. In this case, cognitive burden is posed as a cost that affects fitness, and I will study some properties of behaviour under a pressure to minimise this cost. Again, an important theme is that of the constraints and trade-offs that arise from such a cost. Some examples of such constraints and trade-offs that are studied in behavioural ecology, that are closely related to the ones I will consider, include that of limited attention [30], and the trade-off between speed and accuracy in optimal sampling and signal detection [1].

These results, and the work performed here, show the effects of trade-offs on high level behaviour generation. Physical constraints can however be found all the way down to the earliest stages of perception, governing the maximum resolution of lens and compound eyes [48], and trade-offs between heat dissipation, energy consumption and temporal and spatial resolution in visuo-motor systems [53]. The work in this thesis gathers a range of these specific costs under the title of 'cognitive burden', allowing the study of their consequences in a principled way by applying information-theoretical tools, without the need to formulate their exact source and form explicitly.

Finally, the term 'cognitive burden' often comes up in work on cognitive load theory [72], which studies optimal teaching and learning methods given the premise of limited cognitive capacity of the learner. Although this premise is shared by the work in this thesis, I will make no attempt to extend the results towards recommendations for best practices in the classroom.

Chapter 4

Some Effects of Cognitive Organisation on Bandwidth Requirements

> It is a very sad thing that nowadays there is so little useless information.
>
> *Oscar Wilde*

4.1 Introduction

Information processors are ubiquitous. In the digital age that we currently live in, the first examples you may think of is your PC, or more recently, your smart phone. These are capable of taking some information, process it, and perform some operation based on the outcome of this processing. When multiple of such information processing devices work together, the processing capabilities grow larger and larger, culminating in the vast information processing infrastructure that is the internet.

There is also a class of more traditional information processors that form structures at least as complex: organisms. We, as well as other animals, and even very simple organisms such as bacteria, can be considered as processing information about the world, and acting as a result of this processing. This process can be very intricate, involving a large amount of information. Rather than processing this amount in one central place, again the infrastructure, i.e. our brain, is made up of many smaller information processing units. The structure of these combined units achieve the processing requirements for performing what we call intelligent behaviour.

In this chapter, I will provide an initial survey into this effect: how the structure of the system in which information is processed affects the bandwidth requirements for this processing, both for separate units, and for the system as a whole. I will do this using the model of an agent acting in an artificial environment, but the methods that are used, and the insights derived from applying them, are kept as general as possible, such that they can be applied to a range of information processors that observe and interact with the environment they are in. This is achieved by specialising the generic information theoretic framework, constructed from the methods described in the previous chapter. This framework was agnostic about the actual form and implementation of the information

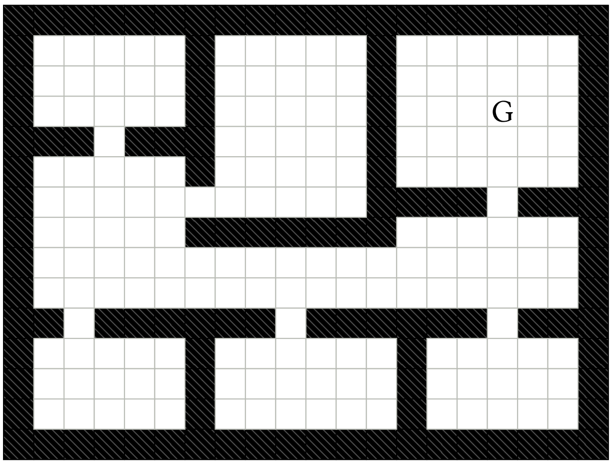


Figure 4.1: Grid world environment used in experiments. The state of the world is described by the cell that the agent occupies. The agent can move north, east, south, or west. Every action has a negative reward of -1, except when that action brings the agent into the goal cell marked with '*G*'.

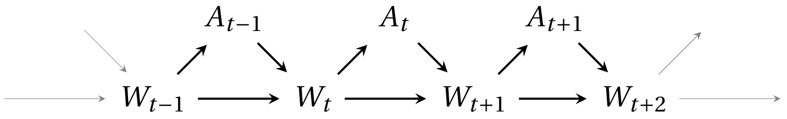


Figure 4.2: Causal Bayesian Network of the perception-action loop. In this case, full observability is assumed, which allows the world and sensor states to be contracted into a single variable W_t.

processing units that it deals with. In this chapter, I will step away from this and introduce some organisation. The results obtained by doing so will give a better understanding of the effects of different aspects of the perception-action loop, and form a guideline for the rest of this thesis.

4.2 Agent and Environment

Consider a grid-world scenario, as shown in 4.1, which is similar to popular navigation problems used in machine learning literature [97, 98, 88]. At any time, an agent is positioned in one of the free grid cells. This position determines the state of the world. The set of possible positions is denoted $\mathscr{W}$. The agent can move around the world, having 4 different actions at its disposal in its action set $\mathscr{A}$: move north, east, south or west. Its goal is to reach the cell marked 'G', in as few steps as possible. This goal is modelled using a reward function that gives a negative reward (i.e. a cost) for each step that does not take the agent to the goal: $R^a_{ww'} = -1$ if $w' \neq g$, 0 otherwise, where g is the state in which the agent is at grid cell 'G'.

As discussed in the previous chapter, if we treat the state at some time t and the action selected by the agent at that time as random variables W_t and A_t, the agent-environment interactions can be formally described by a causal Bayesian network, as reproduced in

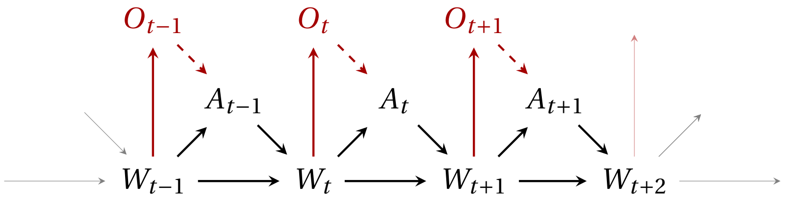


Figure 4.3: Causal Bayesian network of perception-action loop with layered action selection. State information is used to select an initial option O_t. This option, plus additional state information, is then used to select an action. Dashed edges indicate connections that are purely informational, rather than causal.

Fig. 4.2. The relevant information method can then be applied to find the minimum amount of state information that goes into action selection on average at each time step, measured with $I(W_t; A_t)$, to achieve a certain amount of expected utility averaged over all states and actions, as well as the policy that achieves this minimum. In this scenario this amount, the relevant state information, is 1.39 bits in order to achieve maximum performance[1]. In the remainder, I will denote the policy achieving this minimal bandwidth with π_{flat}.

It must be stressed again that this is a fundamental minimum: no reactive policy can be optimal without requiring at least a bandwidth of 1.39 bits. From the discussion in the introduction of this chapter it may seem that this constraint may be alleviated by using a more intricate organisation of the cognitive system than a single action selection unit. However, as a result from the data processing inequality, no structure could possibly reduce the total amount of information that is required. What *may* be possible to achieve, is to spread the information processing over several smaller decision making units, or reduce the average amount of information *intake* by having access to relevant information from another source than direct observation of the world..

To study such effects, I will look at three slightly more complex action selection structures: 1) *layered action selection*, in which action selection is done in stages. 2) *action selection with action memory*, in which a memory of the previous action is available. 3) *dispersed processing with memory*, in which again action selection is done in stages, but with intermediate decisions persisting for a longer time.

4.3 Layered Action Selection

I present a simple model of layered action selection. Instead of selecting an action directly, an initial *option*[2] is chosen first, according to some distribution $p(o_t|w_t)$. Next, the actual action is selected, according to the standard policy $\pi(a_t|w_t)$. Note that the action selection is not directly influenced by the selected option; rather, the option acts as 'side

[1] To calculate the mutual information, here, and for similar measures in the remainder of this chapter, $p(w_t)$ is assumed to be uniform. This choice will be discussed later.

[2] This terminology is taken from the field of hierarchical reinforcement learning, where options are higher level, temporally extended abstract actions[98]. The options of this chapter (will) have some of these properties, but do not implement those of the RL framework precisely.

information', that is accessible by the action selection process. This is indicated in Fig. 4.3 by the dashed edges, which are not purely causal connections, but rather indicate relevance. If one considers O_t to be a bottleneck variable, a view I will arrive at further down, the solid edges form G_{in}, and the dashed edges make up G_{out}.

Having access to this side information may reduce the additional sensory bandwidth needed to select an action: if O_t predicts A_t well, only a small amount of additional world information needs be accessed. However, of course now there are two information channels that may incur an informational cost: one selecting the option, and one selecting the action with the option as side information. The *total* average information use per step is the mutual information between state and the combination of the selected option and action $I(W_t; A_t, O_t)$. This is equal to the sum of the information captured in the option selection and the additional information that is still required given this option: $I(W_t; A_t, O_t) = I(W_t; O_t) + I(W_t; A_t|O_t)$. According to the parsimony hypothesis, there is a drive to minimise this quantity, so the interest is again on the minimal case.

Since it also holds that $I(W_t; A_t, O_t) = I(W_t; A_t) + I(W_t; O_t|A_t)$, and $I(W_t; O_t|A_t) \geq 0$, one can quickly conclude one property of layered action selection: it will never reduce the total amount of information processing. A trivial solution to a naive minimisation of the total over all valid mappings $p(o_t|w_t)$ therefore is one that results in $I(W_t; O_t) = 0$, and $I(W_t; A_t|O_t) = I(W_t; A_t)$. In other words, the layered process is simply flattened.

However, the benefit of adding another layer of pre-processing is to limit the informational burden on the action selection process, so to find a useful mapping we can add a constraint on the information requirements of this part. This results in the following problem:

$$\min_{p(o|w)} I(W_t; O_t, A_t) \quad \text{subj. to} \quad I(W_t; A_t|O_t) \leq C \tag{4.1}$$

With a simple proof (see Sec. 4.7), it can be shown that solving this problem is equivalent to solving:

$$\min_{p(o|w)} I(W_t; O_t) \quad \text{subj. to} \quad I(O_t; A_t) \geq \tilde{C}. \tag{4.2}$$

This is useful since this new problem has the form of a standard information bottleneck, which confirms the view of O_t as a bottleneck variable. This means that one can apply the well-established IB methods. In particular, we can use the iterative Blahut-Arimoto-type IB algorithm to solve the problem, as described in Sec. 2.2.3. As you may recall, this involves finding the zero of the derivative of the Lagrange equation

$$I(W_t; O_t) - \beta I(O_t; A_t), \tag{4.3}$$

where we can trace β through the full range $[0, \infty)$ to find the optimal trade-off curve.

The curve that is found by doing so for the layered action selection scenario in our grid world is shown in Fig. 4.4. In this experiment, the number of options $|\mathcal{O}|$ is set to two, so it effects a real bottleneck. For comparison, the figure also shows the imaginary line that the curve should have followed to keep the total information intake $I(W_t; A_t, O_t)$ at the same level as in the flat case (i.e. 1.39 bit). The actual curve however lies above

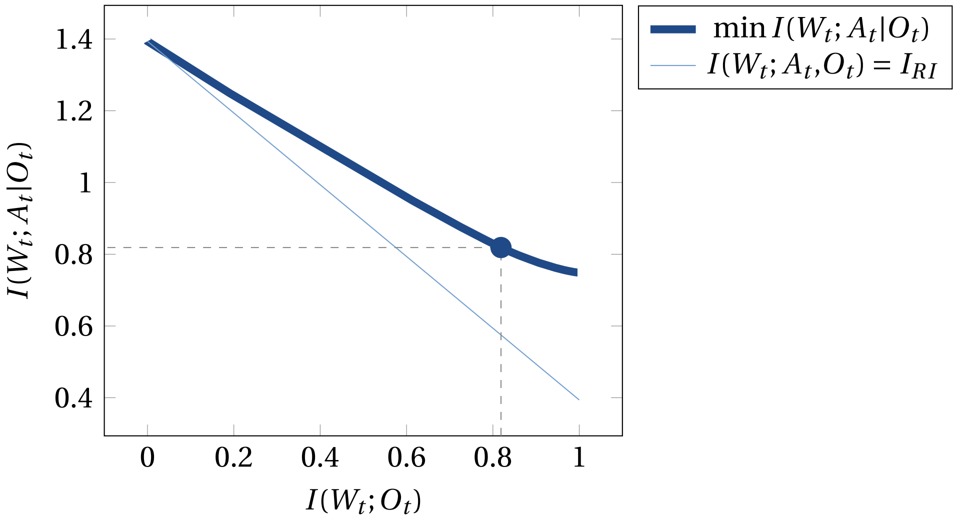


Figure 4.4: Trade-off between state information required at each stage in a two-layered system. Flattening, i.e. no effective use of options occurred at top left ($\beta \to 0$), (near) maximal bandwidth of 1 bit for the initial option selection channel at bottom right ($\beta \to \infty$). Where the required bandwidths of both stages are equal, i.e. where the minimum bandwidth for either of the channels is smallest, found by iterative bisection of the plotted range of β, is marked, at $I(W_t;O_t) = I(W_t;A_t|O_t) = 0.819$, with $\beta = 1.77$. For comparison, the imaginary line that the curve should have followed to keep the total intake at the minimum of the relevant state information is marked with the thin diagonal line.

this line, which means that the layered structure indeed increases the total information requirements. In other words, the information needed at the top level is more than what can be saved at the lower level.

Each stage *individually* however requires less information than in the flat case along the whole curve, except for $\beta = 0$ (leftmost point on the curve, which reduces to the flat framework). An interesting trade-off is marked in the graph with a circle and dashed lines: here the information requirements at each level are equal. Since all points on the trade-off curve are Pareto optimal, this point is the solution to $\min_\beta \max(I(W_t;O_t), I(W_t;A_t|O_t))$. In other words, at this point the maximum bandwidth of the most complex component is minimized.

This point is here found by iterative bisection of the range of β until it becomes sufficiently small. At this point, the units at each individual level need just over 0.8 bit of state information, which is significantly lower than the 1.39 bits that a single unit in a flat structure uses. In other words, even though the total system requires more information, the state information channel of each separate sub part requires a much lower bandwidth.

Figure 4.5 shows the mapping that results from the trade-off solution marked in Fig. 4.4. The first option is chosen with probability 1 whenever the agent should move east, the second whenever it has to move west. The selection becomes stochastic when the agent should move only either north or south. From this we can see how the action selection process, and with it the information intake, is divided over the two layers: the top layer uses some information to make distinctions along the east-west axis, whereas the lower

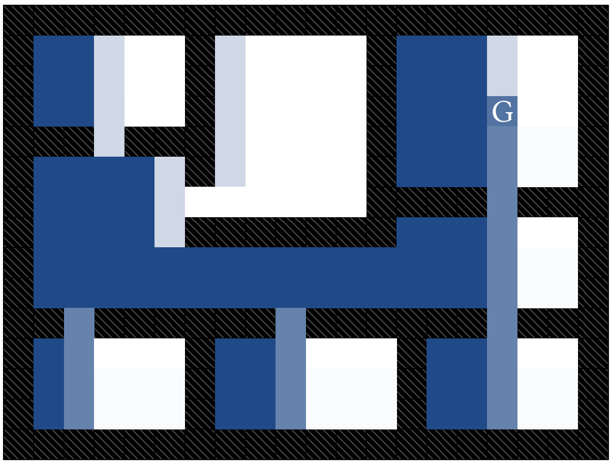


Figure 4.5: Visualisation of option selection in the layered action selection equilibrium marked in Fig. 4.4. Dark shading signifies a high probability of selecting the first option in a cell, light shading shows preference for the other option.

level uses further information to differentiate between north and south. Together they process all information needed to decide on the correct action.

4.4 Action Memory

Next, I investigate a simple model of a memory effect, which can be considered to be orthogonal to that of the previous section: instead of decomposing immediate relevant information, we will see how information stays relevant and can be reused over time. In the flat and layered cases above, the agent was assumed to be highly 'demented': after each action he would forget everything and needs to take in all information that is relevant in the new state. However, it is likely that there is a high degree of correlation between a series of states, due to the way actions affect state and as such feed information back into the world. There can be a large overlap between the relevant information at subsequent steps, which means that due to this redundancy, the total amount of state information that goes into a full action sequence is probably much lower than the sum over time of single step relevant information.

This actual amount of information from one sequence (states) that 'flows' into a second sequence (actions) can be measured with Massey's *directed information* [66]:

$$I(W^N \to A^N) = \sum_{t=1}^{N} I(W^t; A_t | A^{t-1}), \tag{4.4}$$

where the notation X^N denotes a random variable constructed from a sequence of length N of readouts of variable X. The arrow indicates the directionality of this measure: it only measures information flowing from one sequence to the other, not counting any feedback from the output into the input. In other words, it measures how an action sequence is influenced by a state sequence, not how actions influence states. Normal mutual information does not discard this feedback, and thus generally it holds that $I(W^N \to A^N) \leq I(W^N; A^N)$.

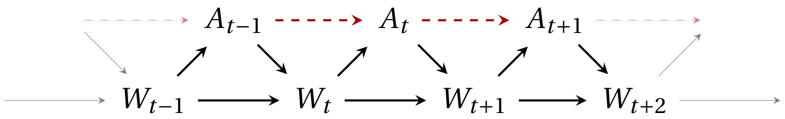


Figure 4.6: Causal Bayesian Network of one-step action memory model. To decide upon an action, the agent now has access to a memory of the action selected in the previous time step. This could reduce the additional state information that needs to be accessed.

The directed information is a measurement on a whole sequence of length N. To acquire a per-step average information measurement that can be compared to those used so far, one can measure the *directed information rate* in the limit of an infinite sequence, $\lim_{n\to\infty} \frac{1}{n} I(W^n \to A^n)$. An iterative method exists to estimate this quantity, while running the process described by the state transition model and agent policy [113]. Applying this method to the current scenario, and the RI-optimal flat policy π_{flat}, we arrive at a rate of 0.064 bit, less than 5% (!) of the total per step relevant information. This means that more than 95% of the information taken in and processed by a memory-less agent is redundant: this information was already captured in previous states and/or actions. An agent with complete, or in any case sufficient memory of the past could therefore reduce its sensory bandwidth requirements by a factor of 20.

A complete memory of the past however may be unrealistic, and the required memory bandwidth may outweigh the reduced cost on the sensory bandwidth. Therefore, it is instructive to also study the most simple memory model: where the agent only remembers its previous action. This is modelled with the CBN of Fig. 4.6. Remember that the dashed edges indicate side information, not strictly causal influences. In this scenario, the state information intake of the agent is described by the 2-step horizon directed information:

$$I(W^2 \to A^2) = I(W_1; A_1) + I(W_1, W_2; A_2 | A_1). \tag{4.5}$$

The first term in the sum on the right is the average immediate information intake of an agent dropped in the world at a random state without any memory, and the second term, which due to Markovianess reduces to $I(W_2; A_2|A_1)$, gives the average additional information needed, given a memory of the previous action, in subsequent steps. Determining this term in our grid-world, again with the optimal flat policy of before, shows that this intake drops to 1.14 bits. So, here even a limited action memory of just one step already provides 20% of the relevant information.

But perhaps the agent can do even better than that. Maybe it could shape its policy in such a way that its previous actions become more informative. It could for instance use a rule such as ‘move east after every time I moved south’. Each time it applies this rule, it does not need to observe any state information directly. In other words, we want to minimise the second term, the information intake given the knowledge of the previous action, over all policies, to a given amount. This can only be done by increasing the information intake at the previous step, which results in the trade-off formulated as the following minimisation problem:

$$\min_{\pi(a|w)\in\Pi^*} I(W_t; A_t) \quad \text{subj. to} \quad I(W_{t+1}; A_{t+1}|A_t) \leq C, \tag{4.6}$$

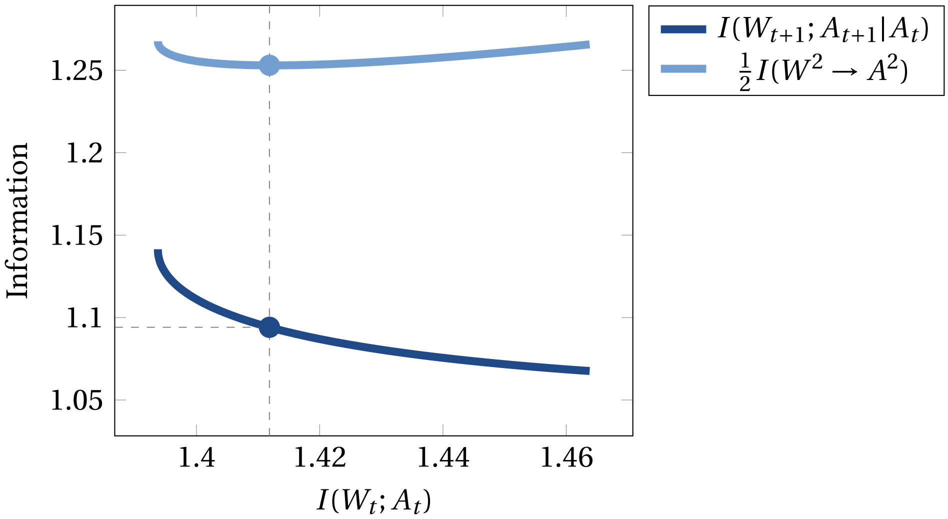


Figure 4.7: Dark curve: trade-off between state information required at subsequent steps in an action memory model. The lighter curve shows the resulting average over two steps, or the normalised two-step directed information. The minimum of this curve is marked, which gives the optimal average information intake for an agent that remembers its last action every second step.

where Π^* is the set of all policies that are optimal in the sense of expected utility. Again it can be shown (c.q App. 4.7.1) that this problem is equivalent to one similar to an IB normal form, namely:

$$\min_{p(a|w)\in\Pi^*} I(W_t; A_t) \quad \text{subj. to} \quad I(A_t; A_{t+1}) \geq \tilde{C}, \tag{4.7}$$

This means that as before, the standard IB methods can be used to trace out the full trade-off curve, which is shown as the dark line in Fig. 4.7. It shows that indeed it is possible to pick a policy that can benefit even more from a single step action memory, in terms of additional information requirements, up to the point where the conditional information drops to 1.07 bits, just over 75% of the memory-less case.

This reduction comes at the cost of an increase in information intake at the initial step, to 1.46 bits at the end of the curve. Over a long period this initial intake may be negligible. However, if the agent loses its action memory frequently, perhaps because its limited memory was needed to store something else, or if the previous action often becomes irrelevant due to noise, these additional memory ‘bootstrapping’ costs may outweigh the benefits. If we consider for instance the extreme example where the agent has no memory every second step, the average intake becomes the normalised two-step directed information, $\frac{1}{2}I(W^2 \to A^2)$. The light curve in Fig. 4.7 shows this average for the trade-offs found on the dark curve, and we see indeed that there is a minimum, at $I(W_t; A_t) = 1.41$ and $I(W_{t+1}; A_{t+1}|A_t) = 1.09$, after which the initial intake grows faster than the subsequent savings.

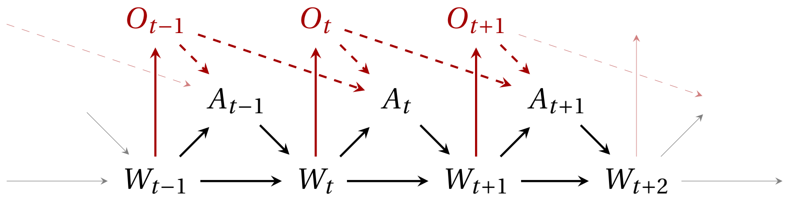


Figure 4.8: Causal Bayesian Network of layered action selection with memory. Here, at any time step the agent has access to both the option currently selected, as to a memory of the previously selected option.

4.5 Layered Action Selection with Memory

Finally, we will look at a combination of the effects of the action selection organisation of the previous two sections: layered decision making, and memory. To do so, the layered model can be extended, to give the agent a memory of the option selected at the previous time-step. The new CBN corresponding to this model is shown in Fig. 4.8.

There are now three possible sources of information relevant to action selection: the state W_t as always, the initial decision of this time step O_t, and that of the previous time step, O_{t-1}. Note that there is no direct informational connection assumed between the previous action and the next: information about the previous action that may be relevant can now be more explicitly stored in the option choice.

If we again postulate the aim of the option selection to decrease the necessary further sensory input, but this time not only for this time step, but also for the next, we arrive at the new problem:

$$\min_{p(o|w)} I(W_t;O_t,A_t) \quad \text{subj. to} \quad I(W_t;A_t|O_t) \leq C_1, I(W_{t+1};A_{t+1}|O_t) \leq C_2 \tag{4.8}$$

To be fully correct, the first term of the constraint should also be conditioned on O_{t-1}, as the value of this variable is known to the agent at t; adding it would capture the desire to have the previous decision carry relevant information for the current action. This would however make the solution much more complex by introducing a double dependency on $p(o|w)$ in the constraint. Also, due to stationarity, the second term already forces options to be informative for the next action, so for simplicity this is omitted here. In the next chapter I will develop methods that deal with future informational effects of current decisions more rigorously.

Once again, this problem is equivalent to the following (multivariate) IB formulation (c.q. App. 4.7.1):

$$\min_{p(o|w)} I(W_t;O_t) \quad \text{subj. to} \quad I(O_t;A_t) \geq \tilde{C}_1, I(O_t;A_{t+1}) \geq \tilde{C}_2, \tag{4.9}$$

leading to the Lagrange equation:

$$L\big(p(o|w),\beta\big) = I(W_t;O_t) - \beta_1 I(O_t;A_t) - \beta_2 I(O_t;A_{t+1}). \tag{4.10}$$

If one sets $\beta_1 = \beta_2 = \beta$, again we can find the zero of this equation, along the range $\beta \in [0,\infty]$, which we will do for our scenario. First, however, a base policy $\pi(a|w)$ needs

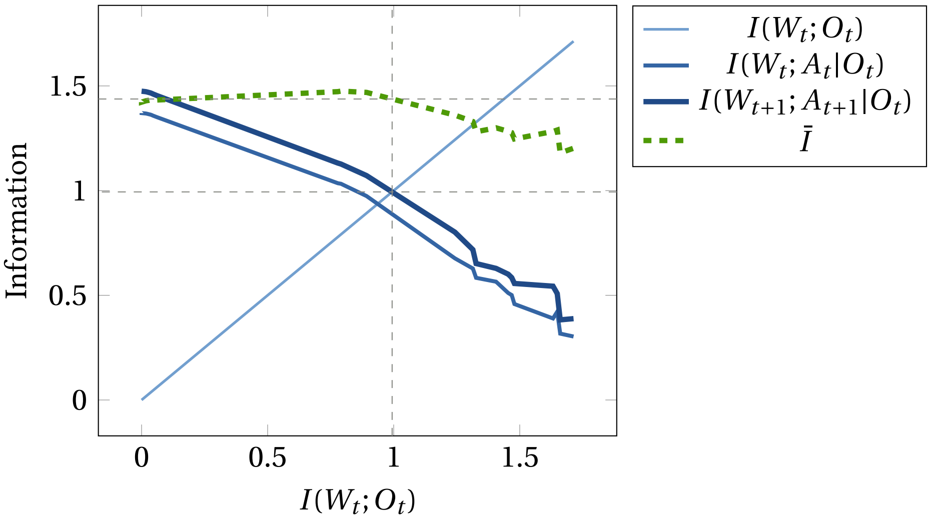


Figure 4.9: Informational trade-offs found for layered decision making with memory. The solid lines give the information bandwidth of the system's parts: option selection (thin light line), immediate action selection given option (medium line), and second step action selection given option (thick dark line). The dashed line shows the average per-step intake, $\bar{I} = \frac{1}{2}[I(W_t;O_t)+I(W_t;A_t|O_t)+I(W_{t+1};A_{t+1}|O_t)]$. The point where the bandwidth of each part is equal, and the maximum bandwidth over parts is minimised, is marked with thin dashed lines. The lack of smoothness at the right end of the graphs is caused to limitations of the IB algorithm; see the main text for details.

to be chosen again; for the results given below, the policy that minimises the normalised two-step directed information of the previous section is used, as we know that it has a good potential for capturing current information that is relevant in the future. For these experiments, the cardinality of the option set is doubled, to $|\mathcal{O}| := 4$, to give the option selection channel enough potential bandwidth to capture information relevant in two time steps.

The results of finding the trade-offs in this model are given in Fig. 4.9. Trade-offs are shown for $\beta \in [0 - 0.25]$, for higher values the iterative IB algorithm was not able to find valid trade-offs, even after 100 iterations. The difficulty of finding good global optima is visible in the lack of smoothness at the high end of the curves. Performing more iterations would smoothen the graph further, however the selected amount and range for β is already sufficient to determine the general trends and some interesting trade-offs.

As expected, the low level bandwidths for both the current step and the next step drop as the higher level decision captures more and more information, and rather significantly so. The dashed line shows that for low β, the information intake at the higher level balances the drop in bandwidth at the lower level. For higher beta, when more than 1 bit is processed by the higher level, we see that the total average per-step bandwidth *does* drop, as it did in the action-memory model. In contrast with that model however, there is a range where this total is lower *and* the individual bandwidth of each part is lower than required in the flat case. This shows, that with the correct organisation of processing units, it *is* possible to reduce the sensory bandwidth requirements of the whole system,

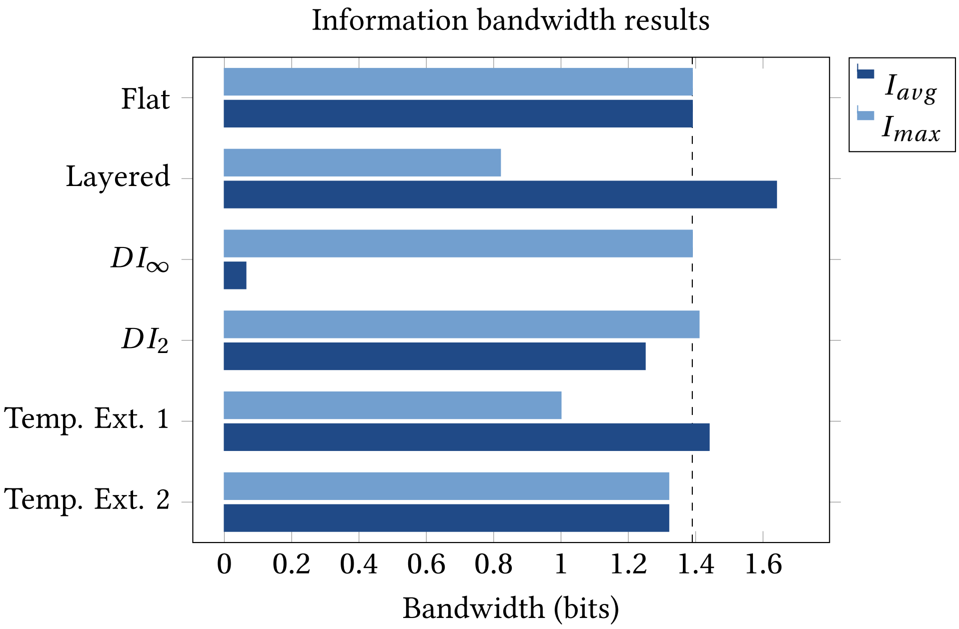


Figure 4.10: Summary of bandwidth results shown for the average per-step bandwidth of the full system (I_{avg}), and for the maximum average for a single component of the architecture (I_{max}), for different models: flat, layered, infinite directed information rate (= infinite action memory), normalised 2-step directed information (= 1-step action memory with 50% memory loss), and two for layered selection with memory; 1: trade-off that minimises I_{max}, and 2: trade-off where $I_{max} = I_{avg}$.

and of each separate unit at the same time.

The thin dashed lines mark the trade-offs where the maximum bandwidth required for a single unit is minimised. The green line shows that here the total per step bandwidth is similar to the flat case.

4.6 Discussion

In this chapter, I have looked at how some different organisations of a decision making and acting system can affect the informational bandwidth requirements of that system and its parts. A summary of the bandwidth requirements for the studied frameworks are show in Fig. 4.10. From these results, one can make several observations, some of which form the foundation of the upcoming work in this thesis.

4.6.1 Dividing decision making over multiple layers can significantly decrease the bandwidth at each separate layer.

The global decrease however comes at the cost of a higher bandwidth of the total system. This result is not specific to our scenario: stemming from the data processing inequality, it can be shown that indeed the total bandwidth will *never* drop due to layered decision making. Moreover, experience has shown that it is common that in an information bottleneck, the information between input and bottleneck variable is not just non-less, but *significantly larger* than the information that is ultimately encoded in the output variable.

This will be a recurring theme in this thesis: to be able to be more parsimonious with information somewhere, you will have to pay with a higher bandwidth somewhere else. However, I will also show, and have actually already shown in this chapter, that certain organisations of an agent's cognitive framework and/or its actions can make the benefits outweigh the costs (see below, and chapter 5), that informational costs of one source can be exchanged for that of another (see chapter 6), and that the additional, concomitant information in an information bottleneck may be a driving evolutionary force (see chapter 7).

4.6.2 Memory can reduce overall bandwidth, due to capturing temporal redundancy.

The infinite directed information rate results show that a memory of all previous actions can greatly reduce the amount of additional information that is still required. It is important to note however that the dramatic drop in our scenario is partly due to the fact that the world is fully deterministic, i.e. performing the same action in the same state always has the same result. Because of this, an agent that takes in all the available state information when it is placed in the world, can thereafter blindly move to the goal state along an optimal path; the start state combined with the action history fully predicts the current state. In a less deterministic world, previous actions become less and less relevant, the further in the past they are. However, one can still expect that in anything but the most antagonistically random and unstructured environment, knowledge of the previous action helps in predicting the current state, which is relevant to selecting the next action.

The 2-step DI result shows the same no-free-lunch principle as seen in the layered framework, though inverted: the overall average bandwidth is lowered, but at the cost of a higher immediate bandwidth requirement on memory loss.

4.6.3 Combining layered decision making with memory can result in a free lunch.

The last two rows of Fig. 4.10 shows that layered action selection with memory combines the benefits of the other models: a lower total average bandwidth is achieved, as with action-memory, *as well* as a lower per-unit bandwidth. In other words: more parsimonious units can be organised into a more parsimonious structure than a flat, memoryless decision maker can achieve!

Two trade-off results are shown, one at the trade-off marked with dashed lines in Fig. 4.9, the second at the trade-off where the $I(W_t; O_t)$ curve crosses the green average curve. The second trade-off definitely shows the free-lunch property (though arguably not significantly), but for the first I_{avg} still seems higher than that of the flat case. However, Fig. 4.9 shows that the average at this trade-off *is* nearly equal to that when the higher layer captures no information, i.e. at $I(W_t; O_t) = 0$, which should reduce to the flat case. Where does this discrepancy come from?

One sees that the higher average is caused by a bandwidth measurement for the next step that is higher than that for the current step. The fact that there is a difference is explained by the fact that this next bandwidth, $I(W_{t+1};A_{t+1}|O_t)$ (which equals $I(W_{t+1};A_{t+1})$ because O_t carries no information), is calculated using $p(w_{t+1}) = \sum_{w_t} p(w_{t+1}|w_t)p(w_t)$. This causes generally that $p(w_{t+1}) \neq p(w_t) : w_{t+1} = w_t$, and $I(W_t;A_t) \neq I(W_{t+1};A_{t+1})$. This effect, that the mutual information between impulse and response can be manipulated by choosing to present the impulses with a different distribution, have lead some to argue that mutual information is the wrong measurement for information processing, if the choice can be arbitrary [90]. Luckily, as we will see later, we *can* find the *actual* state distribution in our abstract framework that is consistent with the MDP description and policy. Moreover, the insight that the state distribution depends on the selected policy forms the foundation of the next chapter, which allows me to find new trade-offs and interesting policies for a class of agents.

4.7 Proofs and Derivations

4.7.1 IB Equivalence

Layered

Theorem 1. *Given the CBN of Fig. 4.3, the constrained minimisation problem:*

$$\min_{p(o|w)} I(W_t;O_t,A_t) \quad \textit{subj. to} \quad I(W_t;A_t|O_t) \leq C \tag{4.11}$$

is equivalent to the Information Bottleneck:

$$\min_{p(o|w)} I(W_t;O_t) \quad \textit{subj. to} \quad I(O_t;A_t) \geq C \tag{4.12}$$

Proof. If we make the problem unconstrained using a Lagrange multiplier, we have:

$$\begin{aligned} L(p(o_t|w_t),\beta) &= I(W_t;O_t,A_t) + \beta I(W_t;A_t|O_t) \\ &= I(W_t;O_t) + I(W_t;A_t|O_t) + \beta I(W_t;A_t|O_t) \\ &= I(W_t;O_t) + (1+\beta)\left[I(W_t;A_t) + I(O_t;A_t|W_t) - I(A_t;O_t)\right] \end{aligned}$$

From the CBN, we can see that the Markov property $O_t \rightarrow W_t \rightarrow A_t$ holds, and thus that $I(O_t;A_t|W_t) = 0$. Finally, because $I(W_t;A_t)$ is constant under $p(o_t|w_t)$, the solution of (4.11) is characterised by:

$$\operatorname*{arg\,min}_{p(o_t|w_t)} L(p(o_t|w_t),\beta) = \operatorname*{arg\,min}_{p(o_t|w_t)} \left[I(W_t;O_t) - (1+\beta)I(A_t;O_t)\right] \tag{4.13}$$

Similarly, the solution of (4.12) is characterised by [104]:

$$\operatorname*{arg\,min}_{p(o_t|w_t)} L'(p(o_t|w_t),\beta') = \operatorname*{arg\,min}_{p(o_t|w_t)} \left[I(W_t;O_t) - \beta' I(A_t;O_t)\right],$$

which is equivalent to (4.13), under $\beta = \beta' - 1$. □

Action Memory

Theorem 2. *Given the CBN of Fig. 4.6, the constrained minimisation problem:*

$$\min_{\pi(a|w)} I(W_t;A_t) \quad \textit{subj. to} \quad I(W_{t+1};A_{t+1}|A_t) \le C \tag{4.14}$$

is equivalent to the Information Bottleneck:

$$\min_{\pi(a|w)} I(W_t;A_t) \quad \textit{subj. to} \quad I(A_t;A_{t+1}) \ge C \tag{4.15}$$

Proof. If we make the problem unconstrained using a Lagrange multiplier, we have:

$$\begin{aligned} L(\pi(a_t|w_t),\beta) &= I(W_t;A_t) + \beta I(W_{t+1};A_{t+1}|A_t) \\ &= I(W_t;A_t) + \beta\left[I(W_{t+1},A_t;A_{t+1}) - I(A_t;A_{t+1})\right] \\ &= I(W_t;A_t) + \beta\left[I(W_{t+1};A_{t+1}) + I(A_t;A_{t+1}|W_{t+1}) - I(A_t;A_{t+1})\right] \\ &= (1+\beta)I(W_t;A_t) + \beta\left[I(A_t;A_{t+1}|W_{t+1}) - I(A_t;A_{t+1})\right] \end{aligned} \tag{4.16}$$

From the CBN, we can see that the Markov property $A_t \to W_{t+1} \to A_{t+1}$ holds, and thus that $I(A_t;A_{t+1}|W_{t+1}) = 0$. Thus, the solution of (4.14) is characterised by:

$$\operatorname*{arg\,min}_{p(a_t|w_t)} L(p(a_t|w_t),\beta) = \operatorname*{arg\,min}_{p(a_t|w_t)} \left[(1+\beta)I(W_t;A_t) - \beta I(A_t;A_{t+1})\right] \tag{4.17}$$

Similarly, the solution of (4.15) is characterised by [104]:

$$\operatorname*{arg\,min}_{p(a_t|w_t)} L'(p(a_t|w_t),\beta') = \operatorname*{arg\,min}_{p(a_t|w_t)} \left[I(W_t;A_t) - \beta' I(A_t;A_{t+1})\right],$$

which is equivalent to (4.13), under $\beta' = \frac{\beta}{\beta+1}$. □

Layered Action Selection with Memory

Theorem 3. *Given the CBN of Fig. 4.8, the constrained minimisation problem:*

$$\min_{p(o|w)} I(W_t;O_t,A_t) \quad \textit{subj. to} \quad I(W_t;A_t|O_t) \le C_1 \quad \textit{and} \quad I(W_{t+1};A_{t+1}|O_t) \le C_2 \tag{4.18}$$

is equivalent to the (multivariate) Information Bottleneck:

$$\min_{p(o|w)} I(W_t;O_t) \quad \textit{subj. to} \quad I(O_t;A_t) \ge C_1' \quad \textit{and} \quad I(O_t;A_{t+1}) \ge C_2' \tag{4.19}$$

Proof. If we make the problem unconstrained using Lagrange multipliers, we have:

$$\begin{aligned} L(p(o_t|w_t),\beta_1,\beta_2) &= I(W_t;O_t,A_t) + \beta_1 I(W_t;A_t|O_t) + \beta_2 I(W_{t+1};A_{t+1}|O_t) \\ &= I(W_t;O_t) + I(W_t;A_t|O_t) + \beta_1 I(W_t;A_t|O_t) + \beta_2\left[I(W_{t+1},O_t;A_{t+1}) - I(O_t;A_{t+1})\right] \\ &= I(W_t;O_t) + \\ &\quad (1+\beta_1)\left[I(W_t,O_t;A_t) - I(O_t;A_t)\right] + \\ &\quad \beta_2\left[I(W_{t+1},O_t;A_{t+1}) - I(O_t;A_{t+1})\right] \end{aligned} \tag{4.20}$$

We further know that:

$$I(W_t, O_t; A_t) = I(O_t; A_t|W_t) + I(W_t; A_t) \tag{4.21}$$

$$I(W_{t+1}, O_t; A_{t+1}) = I(O_t; A_{t+1}|W_{t+1}) + I(W_{t+1}; A_{t+1}). \tag{4.22}$$

From the CBN, we can see that the following Markov properties hold: $O_t \to W_t \to A_t$ and $O_t \to W_{t+1} \to A_{t+1}$. Given this, and the fact that $I(W_t; A_t) = I(W_{t+1}; A_{t+1})$ is independent of $p(o|w)$, the solution of (4.18) is characterised by:

$$\begin{aligned}&\underset{p(o_t|w_t)}{\arg\min} L(p(o_t|w_t), \beta_1, \beta_2) = \\ &\underset{p(o_t|w_t)}{\arg\min} \left[I(W_t; O_t) - (1 + \beta_1)(I(O_t; A_t) - \beta_2 I(O_t; A_{t+1})\right]\end{aligned} \tag{4.23}$$

Similarly, the solution of (4.19) is characterised by [92]:

$$\underset{p(o_t|w_t)}{\arg\min} L'(p(o_t|w_t), \beta_1', \beta_2') = \underset{p(a_t|w_t)}{\arg\min} \left[I(W_t; O_t) - \beta_1' I(O_t; A_t) - \beta_2' I(O_t; A_{t+1})\right]$$

which is equivalent to (4.23), under $\beta_1' = 1 + \beta_1$ and $\beta_2' = \beta_2$. □

4.7.2 Self-Consistent Solutions

In this section I show the self-consistent solutions of the three optimisation problems of above, which are used in the iterative methods to solve them, as explained in the main text.

Lemma 1. *The self-consistent solution for the layered system is given by:*

$$p(o_t|w_t) = \frac{1}{\mathcal{Z}} p(o_t) \exp\left[-\beta D_{KL}\big(p(a_t|w_t) \parallel p(a_t|o_t)\big)\right] \tag{4.24}$$

Proof. This is the standard solution for a single-variate information bottleneck, see Tishby et al. [104] for a derivation. □

Lemma 2. *The self-consistent solution for the system with action memory is given by:*

$$p(a_t|w_t) = \frac{1}{\mathcal{Z}} p(a_t) \exp\left[-\beta \sum_{w_{t+1}} p(w_{t+1}|w_t, a_t) D_{KL}\big(p(a_{t+1}|w_t) \parallel p(a_{t+1}|a_t)\big)\right] \tag{4.25}$$

Proof. We seek the solution of:

$$\frac{\delta}{\delta \pi(a_t|w_t)} \left[I(W_t; A_t) - \beta I(A_t; A_{t+1})\right] = 0 \tag{4.26}$$

The derivative of the first term is known to be (c.q. A.2.2):

$$p(w_t) \log \frac{\pi(a_t|w_t)}{p(a_t)} \tag{4.27}$$

Due to the Markov property $A_t \to W_t \to A_{t+1}$, one finds:

$$\frac{\delta p(a_t|a_{t+1})}{\delta \pi(a_t|w_t)} = p(w_t|a_{t+1}) \quad , \qquad \frac{\delta p(a_t)}{\delta p \pi(a_t|w_t)} = p(w_t) \tag{4.28}$$

Filling this in to the derivative of the second term of (4.26), one gets:

$$p(w_t) \log \frac{\pi(a_t|w_t)}{p(a_t)} - \beta p(w_t) \sum_{a_{t+1}} p(a_{t+1}|w_t) \log \frac{p(a_{t+1}|a_t)}{p(a_{t+1})} = 0 \tag{4.29}$$

Finally, rearrangement of terms and adding $p(w_t) \sum_{a_{t+1}} \log \frac{p(a_{t+1}|w_t)}{p(a_t)}$ to both sides of the equality results in the solution of (4.25). □

Lemma 3. *The self-consistent solution for the layered system with memory is:*

$$p(o_t|w_t) = \frac{1}{\mathcal{Z}} p(o_t) \exp\left[-\beta_1 D_{KL}\big(p(a_t|o_t) \parallel p(a_t)\big) - \beta_2 D_{KL}\big(p(a_{t+1}|o_t) \parallel p(a_{t+1})\big)\right] \tag{4.30}$$

Proof. This is the standard solution for a multivariate information bottleneck, with different valuations of the information relevant to different variables; see Slonim at al. [92] for a derivation. □

Chapter 5

Look-Ahead Relevant Information

> “It is always wise to look ahead, but difficult to look further than you can see.”
>
> *Winston Churchill*

5.1 Introduction

I have pointed out an implied assumption of the relevant information method as it is originally formulated, and as applied in the previous chapter: the agent has no better guess for the state distribution than a uniform distribution; it is not able to assess the influence of its actions with regards to this distribution, which clearly is large. As the model of the perception-action loop showed, actions affect the state of the world. Different actions have different results, and choosing one action over the other can change an agent's life course significantly.

In terms of reward, this means that local, immediate decisions determine the long term expected future accumulated reward. It could well be that actions have a higher short term cost, but that this is made up for with a higher long term gain. The knowledge of this is, albeit arguably implicitly, available in the value function. What is missing is a similar consideration for the future effects of actions on informational costs. The following example makes this idea more concrete, and shows how such considerations can change behaviour.

Imagine that you are driving a car, and need to get to the other side of a large city. You are at a crossing with two options: either drive through the centre of the city, or take the motorway around it. It would be no surprise if you prefer the easier second option over having to spend a lot of mental energy on navigating the many crossings in a complex city road layout.

Now imagine that at this initial crossing the option to go through the city is easy: you can simply, blank-mindedly drive straight ahead. In contrast, to get on the orbital road you would need to navigate a complex multi-laned roundabout. How much more extra cognitive load would you be willing to accept locally to be able to avoid all the other difficult crossings and make life easier for yourself in the future? This is analogous to asking how much reward you would be willing to give up now in order to increase future reward. To make it even more interesting: would you be willing to trade performance to make the future easier? In the driving example this could mean still taking the easy orbital, even when that makes the journey take longer.

In this chapter I will consider these kind of questions in more detail. I will do this by extending the relevant information methods to explicitly take into account future effects of actions on informational burden. With this I will show how issues as those from the previous paragraph can be studied formally. Furthermore, this will enable us to step down from abstract off-line solutions, and apply the relevant information framework to on-line learning agents.

5.1.1 Channels with Feedback

Rate-distortion theory is generally applied to channels *without feedback*. This means that a channel's input does not depend on its past output: $p(x_{t+1}|x^t, y^t) = p(x_{t+1}|x^t)$, where x^t and y^t denotes the input and output history up to and including time t. This is the case for instance in data compression, a main application of the theory: the content of the next video frame does not depend on the compression results of the previous frames[1].

It is obvious however that this does not hold when we model an agent's actuation channel in the PA-loop; the world state transitions (and thereby an agent's sensor states) are not solely determined by the previous state, but also by the actions performed by the agent: $p(w_{t+1}|w^t, a^t) = p(w_{t+1}|w_t, a_t) \neq p(w_{t+1}|w_t)$. The main consequence of this is, that when changing an agent's policy, one also changes the marginal distribution over world and sensor states.

Because the classical RI method ignores this feedback, one firstly ignores that the fixed state distribution that is used is most likely not consistent with the final policy, and, more importantly, that local changes in the policy can result in a very different development of the future with very different informational properties.

Channels with feedback have received a great deal of attention throughout the field of information theory, as such channels are abound in practical applications. When making a phone call, we get feedback from the person on the other side about the quality of the phone line, which we can use to choose to repeat information, or to use different wording. An early result by Shannon [87] is that feedback cannot increase the capacity of a memory-less channel, although feedback is still often useful to make coding cheaper. Recently, Kim [47] showed how to determine the capacity of a wide family of feedback channels with memory, using Massey's directed information [66].

Literature on rate-distortion optimisation with feedback is more sparse. Some examples are available, such as an interesting distributed sensing application by Beferull-Lozano et al. [16]. One study that is closely related to the work in this chapter is by Chou and Miao [28], who present a dynamic-programming based approach to achieve optimal rate-distortion trade-offs for streaming media. They however define rate in terms of raw bits of data being transmitted, instead of bits of Shannon information, such that the contribution to the global rate is fixed for each local decision. This is not the case for an agent: the rate cost of choosing an action depends on the global distribution over actions. Another dynamic-programming solution to rate-distortion with feedback is due

[1] Note that some compression methods do depend on past *input*. In this case, $p(x_{t+1}|x^t) \neq p(x_{t+1}|x_t)$, and the channel is said to be one with *memory*.

to Tatikonda [99], who models the selection of an encoder as a control problem. However, his modelling is also not directly transferable to an agent's perception-action loop: firstly his work implies perfect (or at least sufficient) memory of the full history, which I will not require; and secondly, for his purposes he can model feedback as side-information, separate from the input, in contrast to the PA-loop where feedback is inseparably combined in the world state.

5.1.2 Consistent State Distributions

As mentioned earlier, the original relevant information formulation fixes the state distribution to be uniform. In this section I will make an initial extension of the method by explicitly maintaining a state distribution that is consistent with the current policy. This extension would fit an agent that is able to deduce, or learn the distribution resulting from its policy.

Given a fixed policy $\pi(a|w)$, and the transition model $P^a_{ww'}$, the development of the world state is described by a first-order Markov chain, with transition probabilities:

$$p(w'|w) = \sum_a \pi(a|w) p(w'|w,a) \tag{5.1}$$

If all states in an MDP are positive recurrent, which means that the average return time for each state is finite, the process has a *stationary distribution*, which satisfies

$$p(w) = \sum_{w'} p(w'|w) p(w), \tag{5.2}$$

In an episodic scenario the presence of absorbing states invalidates this requirement: after a sufficiently long time the probability of being in such a state approaches 1. To overcome this, we can manipulate $p(w'|w)$ to be uniform (over possible start states) if w is absorbing. The stationary distribution can now be determined, as detailed in App. A.2.1.

Doing this at each iteration of Algo. 1 gives the distribution $p_k(w)$ that we can plug into the next iteration instead of the original fixed $p(w)$, to make the iteration fully consistent. I will refer to this version of the algorithm as *consistent relevant information (C-RI).*

Example

To show how taking into account the consistent state distribution changes outcomes, consider the MDP in Fig. 5.1. The task for the agent is to travel to the goal as quickly as possible. Which policy will an RI optimal agent follow? One may recall from Sec. 2.6, that in information minimisation, the policy in states with the fewest options will shape the policy in the remaining states. In this case, there are 12 states where only one action is optimal; 6 with action 1 (1–5 and 12) and 6 with action 2 (0, 7–9, 11 and 13). So, over these states, the average probability under a uniform distribution for each action is 0.5, which determines the policy for the one state where the agent has a choice: $\pi(a = 1|w = w_6) = \pi(a = 2|w = w_6) = 0.5$.

However, an optimal agent will actually not follow the longer top path; if it starts in w_0, or in the non-episodic case where it is replaced there every time after reaching the goal, the states in this path are never visited at all. This greatly reduces the probability

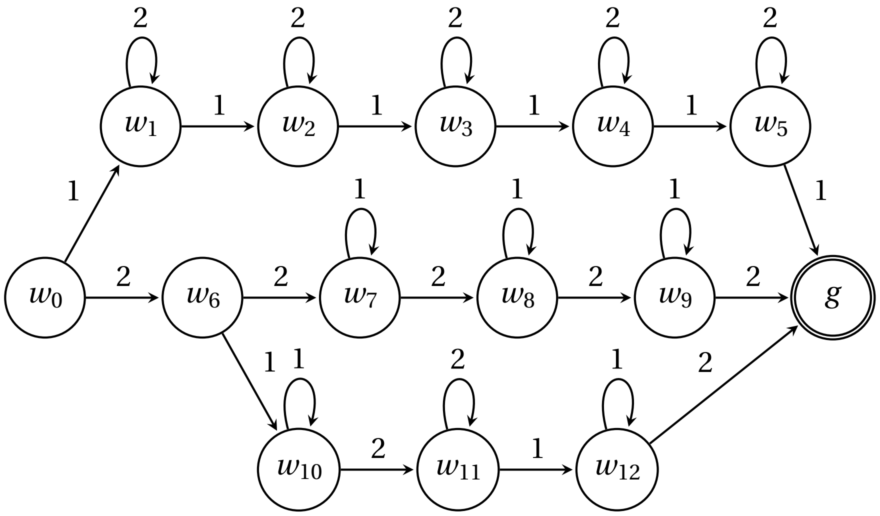


Figure 5.1: Example MDP where taking into account the consistent state distribution results in a different policy. Each step has a cost of 1, except for when the agent enters the goal state, denoted g. This makes only the bottom two paths optimal. The top and middle paths requires choosing the same action (nr. 1 and 2 respectively) at every step, the bottom involves switching between both actions.

that action 1 is chosen, which means that an RI optimal agent has a higher tendency of choosing action 2 when it has a choice, giving it a preference of following the middle path. Ultimately, when applying the RI method with consistent updates, one even settles on a policy where the agent *only* follows the middle path. With this solution, the agent blindly chooses action 2 at every step, reducing $I(W; A)$ to zero.

Counter Example

This method however is not always optimal in all scenarios. Figure 5.2 shows a simple MDP with two possible optimal paths to the goal: one where the agent has to alternate between actions 1 and 2 (top), and one where at each state either action is optimal (bottom). When applying the RI algorithm to this MDP with a uniform state distribution, the solution will be a uniform distribution over actions for all states but w_1-w_4: remember again that the RI method drives the policy towards the average policy over the states with the fewest optimal choices. In this case one can quickly confirm that $I(W; A) = \frac{1}{3}$ bits[2].

There is a much better solution however: the agent should choose action 2 at every time step. This policy is optimal, getting the agent to the goal as fast possible, *and* blind: $I(W; A) = 0$. The C-RI algorithm initialised with this solution does not diverge from it, whereas the original RI algorithm does move away and again converges on the sub-optimal 50-50 solution. When seeding the policy randomly however, neither is able to find the optimal solution; any probability of traversing the upper route is reinforced by pulling the policy in the first state towards being uniform. The algorithms find at every

[2] Note that the start and end states are visited at every run, whereas the other states only in 50% of the runs. So, $p(w_0) = p(g) = \frac{1}{6}$, $p(w_1 \vee w_2 \vee w_3 \vee w_3) = \frac{1}{3}$, and thus $I(W; A) = \frac{1}{3} \log \frac{1}{0.5} = \frac{1}{3}$

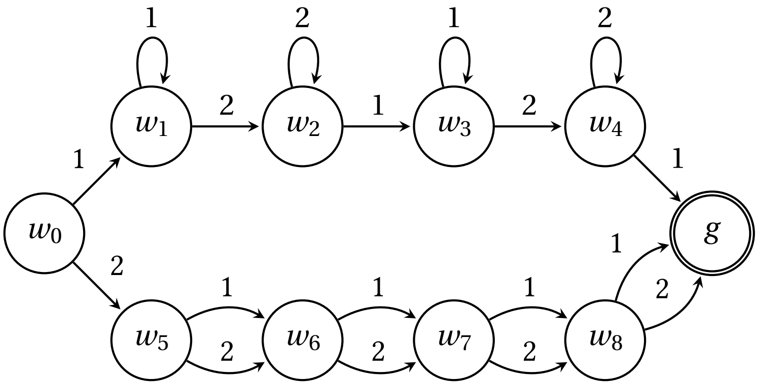


Figure 5.2: Example MDP where taking into account the consistent state distribution is not sufficient to find the most parsimonious policy. Each step has a cost of 1, except for when the agent enters the goal state, denoted g. Both paths are optimal. The top path requires alternately choosing a single action, in the bottom either action can be chosen.

step the best policy for the given state distribution, they do not see that a local policy that is costly against that state distribution can be countered by leading to a path that is less costly in the long run. In the following section I introduce a new RI concept and algorithms that do take such future informational gains into account.

5.2 Look-Ahead Relevant Information

5.2.1 Definition

To develop a method that takes into account the future, we will have to introduce time. Let X_t be the read-out of variable X at *given* time t, and X_T be a read-out at a *random* time step. The relevant information of the previous section did not assume a given time step, and here I will make that explicit by writing $I(W_T; A_T)$. We can now break this quantity down over time.

$$\begin{aligned}
I(W_T; A_T) &= \sum_{w_T} p(w_T) I(W_T = w_T; A_T) \\
&= \sum_{t=0}^{\infty} p(t) \sum_{w_t} p(w_t) I(W_T = w_t; A_T) \\
&= p(t=0) \sum_{w_0} p(w_0) I(W_T = w_0; A_T) + p(t=1) \sum_{w_1} p(w_1) I(W_T = w_1; A_T) + \dots \\
&\overset{1}{=} p(t=0) \sum_{w_0} p(w_0) I(W_T = w_0; A_T) + p(t=0)\gamma \sum_{w_1} p(w_1) I(W_T = w_1; A_T) + \dots \\
&= p(t=0) \left[\sum_{w_0} p(w_0) \right. \\
&\quad \left. \left[I(W_T = w_0; A_T) + \dots + \gamma^k \sum_{w_k} p(w_k | w_0) I(W_T = w_k; A_T) + \dots \right] \right]
\end{aligned} \tag{5.3}$$

For equality 1, it was assumed that the probability of experiencing a time-step depends on the probability of experiencing the previous time-step according to $p(t = k) = \gamma p(t = k-1)$, where γ can be seen as a 'survival' probability, which is the same for each time-step. The final result takes a form similar to a Bellman equation, the recursive part of which I will take out.

Defintion 1. *The* look-ahead information *of a state w_t under policy π, written as $\mathfrak{I}^\pi(w_t)$, is the expected total future action-relevant information intake when starting at that state at time t and subsequently following π:*

$$\mathfrak{I}^\pi(w_t) = I(W_T = w_t; A_T) + \gamma \sum_{w_{t+1}} p(w_{t+1}|w_t)\mathfrak{I}^\pi(w_{t+1}), \tag{5.4}$$

where $p(w_{t+1}|w_t) = \sum_{a_t} \pi(a_t|w_t)p(w_{t+1}|w_t, a_t)$.

With this definition, Eq. (5.3), and the assumption that the process is time homogeneous, we have that:

$$I(W_T; A_T) = p(t = 0) \sum_{w_0} p(w_0)\mathfrak{I}^\pi(w_0). \tag{5.5}$$

If every time-step is as likely, i.e. the future is infinite, $\gamma = 1$, and we get that $I(W_T; A_T) = \lim_{N\to\infty} \frac{1}{N} \sum_{w_0} p(w_0)\mathfrak{I}^\pi(w_0)$. So, as one would expect, the single step information intake is the *look-ahead information rate.*

5.2.2 Information in a Single Experience

Next, we must choose a way for quantifying $I(W_T = w_t; A_T)$, the amount of information taken from a single state instantiation to choose an action. As noted before, this must fulfill $\sum_{w_t} p(w_t)I(W_T = w_t; A_T) = I(W_T; A_T)$. It turns out there is an infinite range of ways to construct a quantity for which this holds, and if one is only interested in the averaged $E[\mathfrak{I}^\pi(W_0)]$, the choice from this range is arbitrary. However, if we also want to analyse the amount of look-ahead information at a specific state, or specific sequences, the choice matters, because here different definitions give different results. As described by DeWeese and Meister [35], out of all the possibilities, there are two definitions that have interesting properties that are of interest in our context, and which are the only definitions that have these properties.

DeWeese and Meister argue that the most important criterion to choose a definition is additivity of the quantity. This means that the amount of information given by two separate symbols should be equal to the information in one, plus the information in the other given that we already know the first: $I(x, y; Z) = I(x; Z) + I(y; Z|x)$. They show that the only definition that satisfies additivity is the reduction of entropy that occurs when observing a symbol: $H(Y) - H(Y|X = x)$.

However, look-ahead information sums over time, and there this definition can give some counter-intuitive results. It is well known that although mutual information is always non-negative, due to $H(Y|X) \leq H(X)$, a specific instantiation of a variable could still increase the entropy of another variable, i.e. it may be that $H(Y|X = x) > H(Y)$. For instance, usually a train may be very punctual, however learning that work is being done on the tracks increases the uncertainty of its arrival time. If this measure is used for look-ahead information, this means that the amount of information at a certain time step becomes negative. If one interprets look-ahead information as the accumulated information intake over time, this would imply that this intake can *drop*, if we increase the horizon over which we measure. Or, put differently, a higher uncertainty about an action than usual at a specific state would 'cancelled out' cognitive burden spent in the past. This does not give a viable definition of information intake and processing over time which fits our intuitions about causality. So instead, what is needed is a formulation where single step information intake is always non-negative.

There is again only a single definition where this holds [35]: the Kullback-Leibler divergence between the posterior and prior distributions of the relevance variable: $I(w_t; A_T) := D_{KL}(p(a_t|w_t)||p(a_t)) = \sum_{a_t} p(a_t|w_t) \log \frac{p(a_t|w_t)}{p(a_t)}$. This is the definition we will use.

5.2.3 New Solution to Trade-Off

Finally, I will develop an algorithm for finding the minimal Look-Ahead information under a fixed performance constraint. This is done in a way similar to the original RI algorithm. First, we construct the same Lagrangian equation as for the RI-algorithm, but this time we fill in Eq. (5.5):

$$\Lambda_{LA}(\pi, \beta) = p(t=0) \sum_{w_0} p(w_0) \mathfrak{I}^{\pi}(w_0) - \beta E[U^{\pi}(W_T, A_T)] \tag{5.6}$$

By taking the partial derivative again this version, equating it to 0 and solving for π, under the assumption that $p(t=0)$ does not depend on the policy (e.g. $p(t=0) = \frac{1}{N}$ when $\gamma \rightarrow 1$), we now arrive at:

$$\pi(a_t|w_t) = \frac{1}{\mathcal{Z}} p(a_t) \exp\left[-\gamma \sum_{w_{t+1}} p(w_{t+1}|w_t, a_t) \mathfrak{I}^{\pi}(w_{t+1}) + \beta U^{\pi}(w_t, a_t) \right] \tag{5.7}$$

Note that this solution is similar to the original single-step RI solution, with the expected future information cost subtracted from the β-weighted expected future reward.

The final algorithm, which I will denote as *LA-RI*, is then obtained by substituting step 6 of the RI algorithm with this equation, and performing Eq. (5.4) as an additional update step at each iteration. Applying this new algorithm until convergence on the problem of Fig. 5.2 gives us the desired result: the optimal policy where the agent deterministically chooses action 2 at each state.

5.2.4 Scenario 1: Avoiding Uncertainty

To display the effects of a drive to minimise the look-ahead information to the relevant minimum, we present an example scenario in the form of a navigation problem. In this scenario, the agent has to traverse the environment, which offers several pathways that the agent can take. Some of these pathways are optimal, while one path results in the agent enduring a higher cost. Also, some paths are deterministic, while other paths are more noisy and require more information to perform the optimal policy. These differences ensure that the agent can increase performance, decrease informational requirements, or

Algorithm 2 Look-Ahead Relevant Information

```
**Require:** $<\mathscr{S},\mathscr{A},P^a_{ss'},R^a_{ss'}>,\beta,\gamma,\epsilon_\pi,\epsilon_U,\epsilon^\pi_\mathfrak{I}$
**Ensure:** $\pi_k = \arg\min_\pi \left[\mathfrak{I}^\pi(S) - \beta E[U(S,A)|\pi]\right]$
$p_0(s) \leftarrow \frac{1}{|\mathscr{W}|}$
$\pi_0(a|s) \leftarrow \frac{1}{|\mathscr{A}|}$
$U_0(s,a) \leftarrow 0$
$\mathfrak{I}^\pi_0(s) \leftarrow 0$
$k = 0$
**repeat**
    $p_k(s) \leftarrow \text{StaticDistribution}(\pi_k,\ P^a_{ss'})$
    $p_k(a) \leftarrow \sum_s p(s)\pi(a|s)$
    $\pi_{k+1}(a|s) \leftarrow \frac{1}{\mathcal{Z}} p_k(a) \exp\left[-\gamma \sum_{w_{t+1}} p(w_{t+1}|w_t,a_t)\mathfrak{I}^\pi(w_{t+1}) + \beta U_k(s,a)\right]$
    $U_{k+1}(s,a) \leftarrow \sum_{w'} P^a_{ss'}\left[R^a_{ss'} + \gamma \sum_{a'} \pi_k(a'|s')U_k(s',a')\right]$
    $\mathfrak{I}^\pi_{k+1}(s) \leftarrow D_{KL}\big(\pi(a|s) \parallel p(a)\big) + \gamma \sum_{a,s'} \pi(a|s)P^a_{ss'}\mathfrak{I}^\pi_k(s')$
    $k = k+1$
**until** $D_{JS}(\pi_k \parallel \pi_{k-1}) \leq \epsilon_\pi$ **and** $\parallel U_k - U_{k-1} \parallel_\infty \leq \epsilon_U$ **and** $\parallel \mathfrak{I}^\pi_k - \mathfrak{I}^\pi_{k-1} \parallel_\infty \leq \epsilon^\pi_\mathfrak{I}$
```

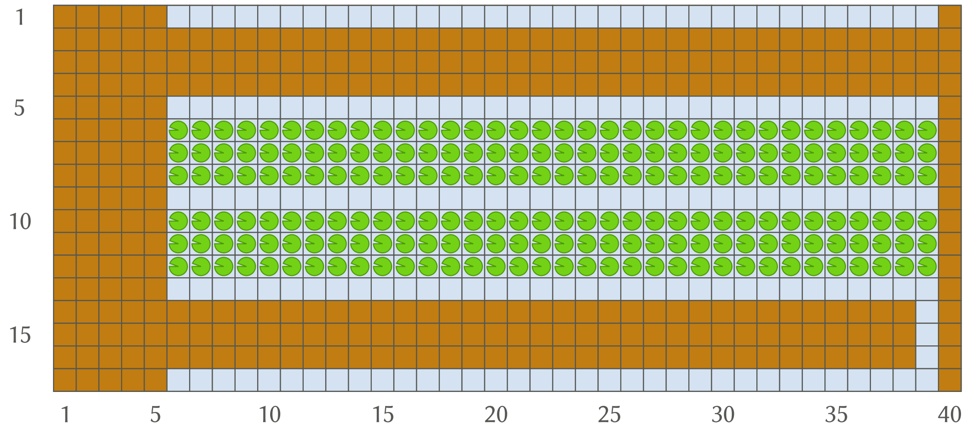


Figure 5.3: The navigation scenario used in the experiments. An agent is placed on one of the west-most cells and has to cross over to the other side of the world, onto one of the east-most cells. Pieces of land are marked with a dark, brown background, and water is denoted with a light, blue shade. The water cells in rows 6-8 and 10-12 are covered by lily pads, denoted with green circles. The agent can move in three ways: jump north-east, east, or south-east. Jumps from land are deterministic, they result in the agent moving in the intended direction. Jumps made from open water are more noisy, 50% of the time these result in the agent landing one cell north or south of where it would normally land. Every jump costs the agent 1 point, however jumping from the water is more difficult, so the agent endures a cost of 10 points when landing into it.

settle on a trade-off, by changing the distribution of the paths it will take to get to the final goal.

More concretely, the agent is placed in a world that consists of a rectangular grid of 40×17 cells, as shown in Fig. 5.3. The world contains a patch of land 5 cells wide in the west and a line of land in the east. There are two land pathways between these pieces of land, in rows 2-4 and 14-16. The southern of the two is cut off by a line of water at the end. Two additional pathways are formed by lily pads, covering rows 6-8 and 10-12. The four pathways are divided by three lines of open water. The state of the world at a given time, S_t, is the location of the agent in the world, which can be in any of the 680 cells in the grid. At each time step the agent can select one of 3 actions: jump either one cell north-east, one cell east, or one cell south-east. The world is bordered by walls in the north and south; performing an action that would have the agent run into the wall results in the agent moving simply eastwards. The world wraps around, such that a move in the last column in the east brings it back to the first column in the west.

Each jump consumes energy, incurring a cost to the agent. This cost is represented by a negative reward: $R^a_{ww'} := -1$. Moving is more costly when the agent has fallen into the water. In this case, the cost goes up to 10, i.e. $R^a_{ww'} := -10$ when in w the agent is in open water. The agent can prevent this by hopping onto the lily pads that float on the water. Finally, the reward is 0 when the agent arrives in one of the goal states, to mark that it has finished the task and to limit the total cost.

A policy thus is already optimal when it brings the agent to the other side, without falling into the water. This is achieved by following the northern land path, or one of the two paths formed by lily pads. However, the lily pads are unstable, making the effect of a jump from one uncertain. With a probability of 0.5, such a jump results in the agent landing either one cell further north or one cell further south than where the action would normally take the agent. The same indeterminacy holds when the agent attempts to jump from open water.

This means that on the two pathways formed by lily pads the agent has to be extra careful not to end up in the water. In fact, on these pathways there always is only a single optimal action available: when next to the open water, try to move away from it, otherwise try to move straight ahead. Any other strategy has the risk of diverting the agent into the water. This means that it has to pay close attention to where it is, to be able to select the correct action.

On the two outer paths, however, the structure of the world offers the agent help to alleviate its cognitive burden. Here, the lack of noise allows it to venture closer to the water and to worry less about which action to take. In each cell on these pathways there are multiple actions that ensure that the agent will not get wet.

In this environment we perform three experiments. Firstly, we determine a policy following the original single-step relevant information method, for different β values, using a uniform state distribution for each iteration. This is Alg. 1, and the method originally introduced by Polani et al [77]. Secondly, we will perform the same experiment, but with the added step of making the state distribution consistent to the current policy at each iteration. To differentiate these experiments, we will refer to the first as the *inconsistent* single-step case, due to the fact that the uniform state distribution that is used generally is not consistent with the policy that is considered. Finally, for the third experiment the

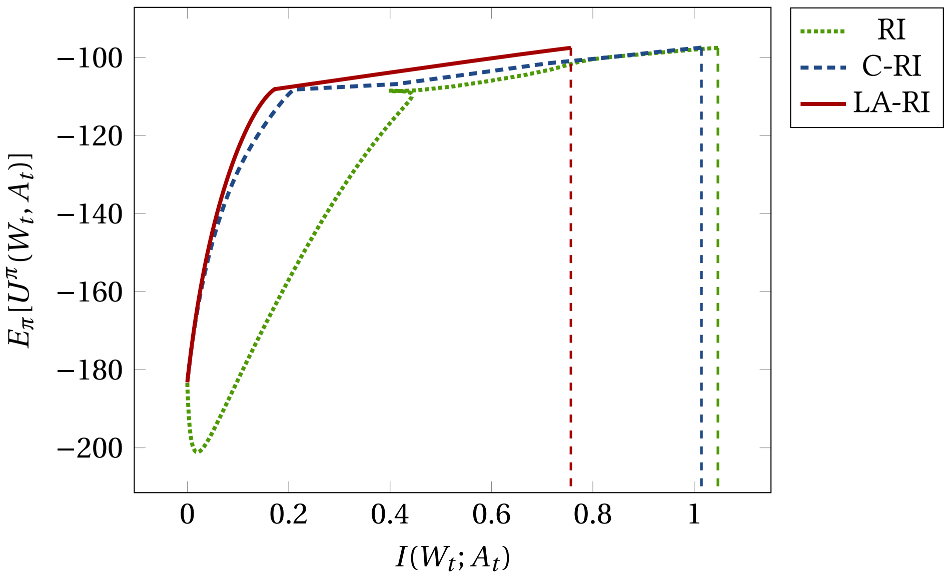


Figure 5.4: Trade-off curves obtained in the navigation task of Fig. 5.3, using the relevant information (RI), relevant information with consistent state distribution (C-RI), and the look-ahead relevant information (LA-RI) methods.

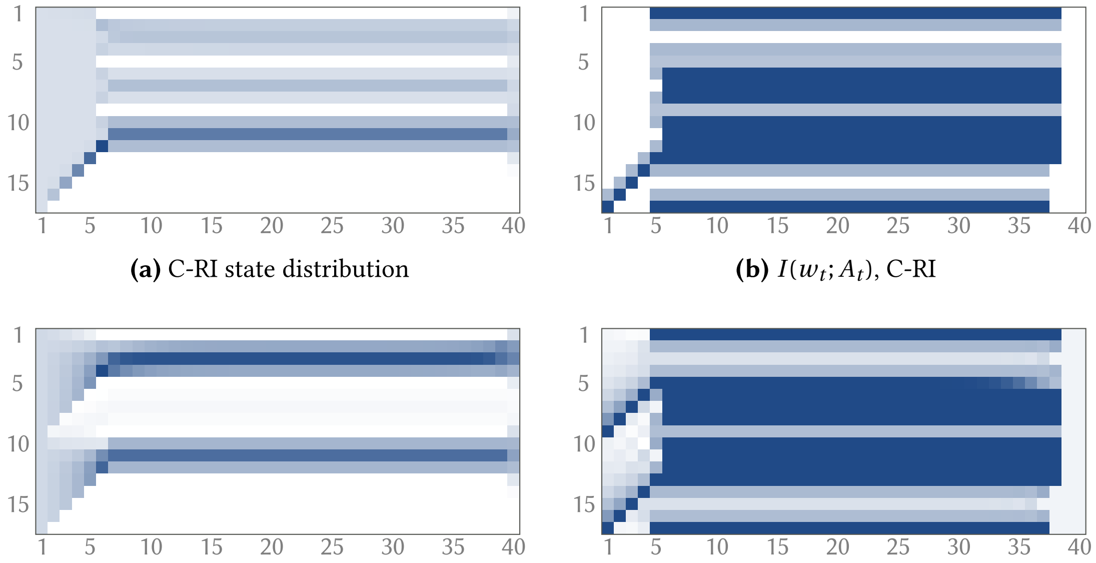
second is repeated using look-ahead information.

For the final policies that are found in these three experiments, we determine the average per-step information intake, $I(W_t; A_t)$, and the performance, measured by the expected utility $E_\pi[U^\pi(W_t, A_t)]$, both using the state distribution *consistent* with the policy. Doing so for the full range of values of β results in the information-performance trade-off curves shown in Fig. 5.4.

The first thing that we notice is that the trade-offs found by the inconsistent relevant information algorithm do not trace out a proper trade-off curve, when considering the consistent state distribution: it is not monotonically growing, and even at some point achieves lower levels of information with rising β. This of course may not be surprising, since the algorithm is not designed to consider changes in the state distribution. The curve for the consistent 1-step algorithm shows that taking these effects into account makes the algorithm converge onto a well-formed trade-off curve, and find policies with a considerably higher performance for the same actual information intake for a significant range.

However, this simple improvement still does not result in finding the true optimal trade-off curve, as is clear from the fact that the curve found by using look-ahead information lies still higher. Especially for high performance, looking ahead resulted in a significant decrease in the information required. At optimal performance, marked by the vertical dashed lines in Fig. 5.4, the decrease is close to 25%, down to 0.76 bit as compared to 1.01 (C-RI) and 1.04 (RI).

We can infer what causes this drop of information from the resulting state distributions, $p(w_t)$. These distributions, shown in Fig. 5.5, indicate how often an agent visits each cell over time. Firstly, we see that both the single-step (a) and the look-ahead (b) methods avoid the lowest pathway, which would force the agent to perform a costly wa-



(a) C-RI state distribution

(b) $I(w_t; A_t)$, C-RI

(c) LA-RI state distribution

(d) $I(w_t; A_t)$, LA-RI

Figure 5.5: Left: State distributions resulting from a) consistent relevant information and c) look-ahead relevant information optimisation, in the environment of Fig. 5.3. Right: immediate information costs per state. Dark shading denotes higher probability/higher information cost.

ter crossing. Secondly, however, the path choice differs on the upper part of the world: whereas the single-step method results in a roughly 50-50 distribution between the top land path and the upper lily pad path, the look-ahead method avoids this second option whenever possible. It is able to see that picking the deterministic topmost path helps to relieve its long term cognitive burden, since it does not have to consider in full detail on which side of the water it is. Moving simply forward is always safe, and having a default option available reduces the sensory bandwidth needs.

5.2.5 Learning Parsimony

The recursive formulation of Def. 1 is very similar to the definition of the value function of Eq. (2.26) that is the main topic of optimisation in Reinforcement Learning. Indeed, the updates of the LA-RI algorithm are reminiscent of that of the *value iteration* algorithm used to find the optimal value function. This leads to the question whether we can take inspiration from the field of RL and apply its methods to informational optimisation.

This entails to learning the relevant methods and which policy is optimal from a set of *experiences*, where each experience consists of the tuple $\langle w_t, a_t, r_t, w_{t+1}\rangle$. Such an experience gives some information about the value function that is to be learned. In *Temporal-Difference learning (TD-learning)* such a tuple is used to determine the accuracy of the current estimate of the value function. This estimate predicts the value of a state to be $V(w_t)$, whereas the value that is observed by the agent is $r_t + \gamma V(w_{t+1})$. The difference between these is the *prediction error*: $\delta = r_t + \gamma V(w_{t+1}) - V(w_t)$. This error can be used to

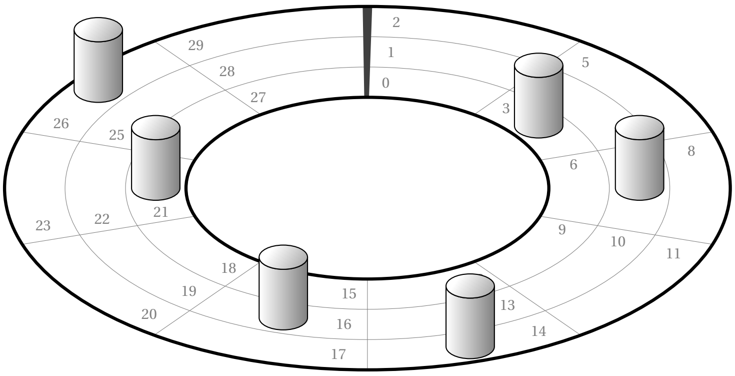


Figure 5.6: Surveillance example. An agent moves clockwise in the environment, choosing to either stay in the same lane, or move one lane out or in. Several obstacles block the path; the agent can move around them, or push them down and step on them, without incurring additional cost.

adjust the value function towards its actual value through the update $V(w_t) \leftarrow V(w_t)+\alpha\delta$, where $\alpha > 0$ is the learning rate.

A similar learning rule can be devised for the LA-information function. Again, the predicted LA-information is $\mathfrak{I}^\pi(w_t)$, whereas the observed amount is $I(w_t;a_t)+\mathfrak{I}^\pi(w_{t+1})$. Here, $I(w_t;a_t) = \log\frac{\pi(a_t|w_t)}{p(a_t)}$, such that $\sum_{a_t}\pi(a_t|w_t)I(w_t;a_t) = I(w_t;A_t)$. The LA-information TD update rule then becomes:

$$\delta = I(w_t;a_t) + \gamma\mathfrak{I}^\pi(w_{t+1}) - \mathfrak{I}^\pi(w_t)$$

$$\mathfrak{I}^\pi(w_t) \leftarrow \mathfrak{I}^\pi(w_t) + \alpha\delta \tag{5.8}$$

Using this update rule, an agent can learn the look-ahead information function on-line solely from experience.

To summarise, we now have the following set of algorithms: original (or consistent) single-step relevant-information ((C)RI, Sec. 2.6), Look-ahead relevant-information value-iteration (LA-RI, Sec. 5.2.3), and look-ahead learning (LA-RIL, current section). In the following section we will apply them to an example problem.

5.2.6 Scenario 2: Clearing Obstacles

The example problem is based on a robotic surveillance scenario. An agent moves around on a circular track, as shown in Fig. 5.6. It takes the agent 10 steps to go around, and there are three lanes, resulting in a total of 30 possible cells the agent can be in. At each step the agent moves one step forward (clock-wise), and he can choose to stay in the same lane, or move one lane left or right; if this would cause him too move outside of the track, he stays in the same track after all.

The agent can not move around freely: there are six obstacles around the track. If a move would bring the agent to a cell blocked by an obstacle, the move is unsuccessful, and the agent remains in the same state. The agent however can also choose to perform a *push* action while moving, which would push an obstacle in the target cell down at the

Parameter	Value
T	10^7
α	$1 - \frac{t}{T}$
γ	0.99
ϵ	$1 - \frac{t}{T}$
β	10^7

Table 5.1: Learning parameters used for the Q-learning and LARIL methods in the surveillance task. α: learning rate, γ: discount factor, ϵ: exploration rate, β: Lagrange parameter (only for RI), T: experiment duration.

same time of the move, making the move possible. Pushed down obstacles come back up again with a probability of 0.1 after each successful round.

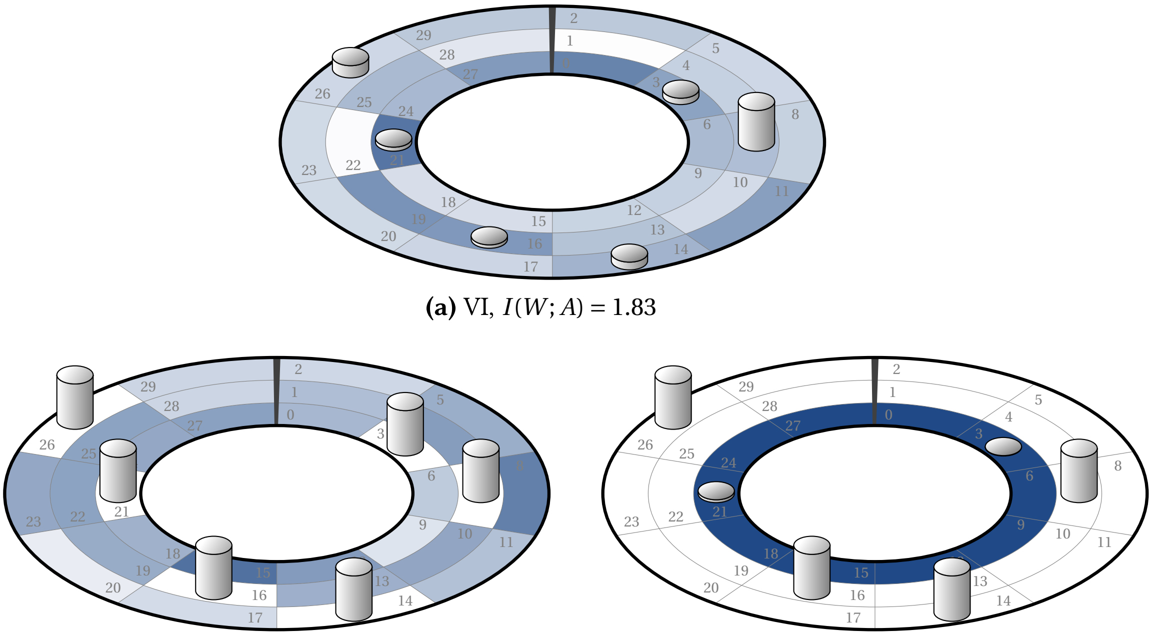
In this scenario, $|\mathscr{S}| = 30 \times 2^6 = 1920$ (number of cells times obstacle configurations) and $|\mathscr{A}| = 6$ (3 moves and whether or not the agent pushes). The reward function, $R^{w_{t+1}}_{w_t,a_t}$ is 1 when the agent finishes a round by crossing the line at the top, -0.1 when the agent performs a push action without there being an obstacle in the target cell (or one that is already down), and 0 otherwise.

There is a whole range of optimal policies in this environment: it is possible for the agent to slalom around the obstacles, but since we did not make pushing down obstacles costly, it may also go around pushing away every one. It is no surprise then that traditional value iteration gives a policy that has the agent moving around the track randomly, pushing obstacles at some steps, moving around them at others. This results in the state distribution as depicted in Fig. 5.7a, and a per-step information requirement of $I(W;A) = 1.83$ bits.

This intake seems excessive, considering that an agent that never pushes only selects from three actions, so should need no more than $\log(3) = 1.58$ bits. Using the RI method we indeed discover a policy that avoids the cells with obstacles, as is evident from the resulting state distribution shown in Fig. 5.7b. It turns out that such a policy even reduces the information intake to less than a quarter of the original.

However, what the single-step RI algorithm does not see, is that an obstacle that is pushed down will remain down for ten rounds on average and that the world is a simpler one during that time. The LA-RI algorithm on the other hand does see this, and with that an even more efficient solution can be found. The policy of this solution moves the agent around in a single lane, only moving straight ahead, pushing down the obstacles that are in the way when they are up. The state distribution of this solution is shown in Fig. 5.7c, and the information intake has been reduced again to less than one third of the RI solution; a full 93% drop from the Q-learning solution.

These results are found using the Blahut/value-iteration type algorithms. Next, we will see if a learning agent is able to find these solutions as well. Firstly, for comparison, an experiment with a Q-learning agent is performed, using the learning parameters as given in Tab. 5.1. Note that the learning and exploration rates decrease linearly over time, as is a popular choice for Q-learning [97]. These parameters are not tweaked to minimise learning time, but are chosen to give the agent a chance to fully explore; decreasing these too



(a) VI, $I(W;A) = 1.83$

(b) C-RI, $I(W;A) = 0.45$

(c) LA-RI, $I(W;A) = 0.13$

Figure 5.7: Surveillance experiments results, depicting the stationary distribution resulting from the policies found using different methods; darker shades indicate higher probability of the agent occupying the given cells, the height of an obstacles is proportional to the probability of that obstacle being up at any random time-step. VI = value iteration, C-RI = consistent single-step relevant information, LA-RI = look-ahead relevant information.

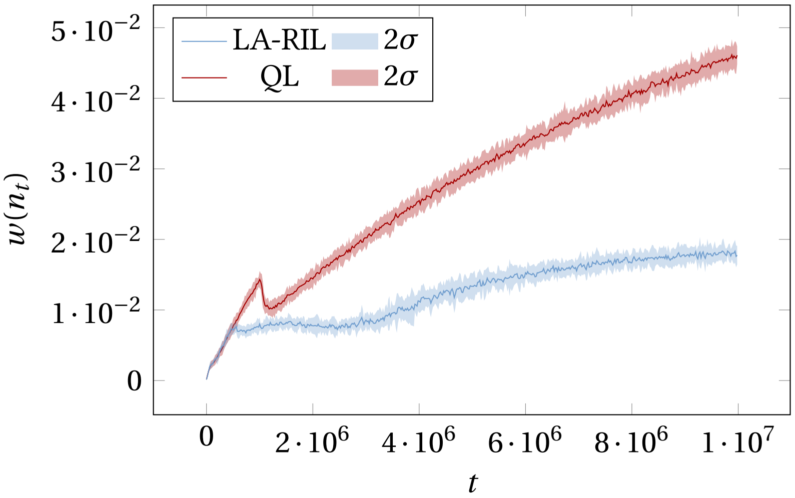


Figure 5.8: The obstacle pushing behaviour of learning agents over time during typical experiments. Successful pushes while the agent is not exploring are recorded as a series of unit impulses. These are then filtered using a sliding Hann window of size 10^5 to get an idea of the development of push frequency over time.

fast prevents the look-ahead information values to be learned accurate enough and makes the agent settle on an informationally sub-optimal policy too quickly, keeping them the high makes the learned value too much based on random actions, instead of those deemed optimal (remember that look-ahead information learning is on-policy) Not surprisingly, the Q-learning agent settles on the same solution as found using value iteration.

Next, we run the same experiment with an agent that performs Look-Ahead Relevant Information learning. During this experiment, the Lagrange trade-off parameter β is set to 10^7, to enforce reward optimal policy. Indeed, the pressure to minimise the look-ahead information does not prevent it from converging on an optimal policy. Moreover, it finds the same RI-optimal policy as the Blahut-type LA-RI algorithm does.

A comparison of the behaviour over time of the two described learning agents is shown in Fig. 5.8. The plain Q-learning agent continuously increases the amount of times it pushes an obstacle down along its learning trajectory. As discussed, this does not increase its performance, but it also does not hurt it. From this agent's point of view pushing or not does not make a difference. At the end of the experiment, it pushes down an obstacle roughly once in every 20 steps, so once every 2 revolutions.

The LA-RI agent on the other hand relatively quickly stops the increase in push frequency in the first part of the experiment; because non-pushing actions are more frequent, pushing an obstacle is against the norm, and uncommon actions increase the information needed to select when to use them. So, the agent avoids directly costly actions and the push frequency remains low. However, as it is gaining more experience, it learns that pushing obstacles out of the way is informationally beneficial in the long run. We see that the push frequency increases again, but still not close to the level of the RL agent. Instead, it only removes obstacles just as often as needed to be able to move around in a simple circle, arriving at the behavior of the LA-RI agent of Fig. 5.7: it removes just the two obstacles in a single lane, which on average results in 2 pushes every 100 steps.

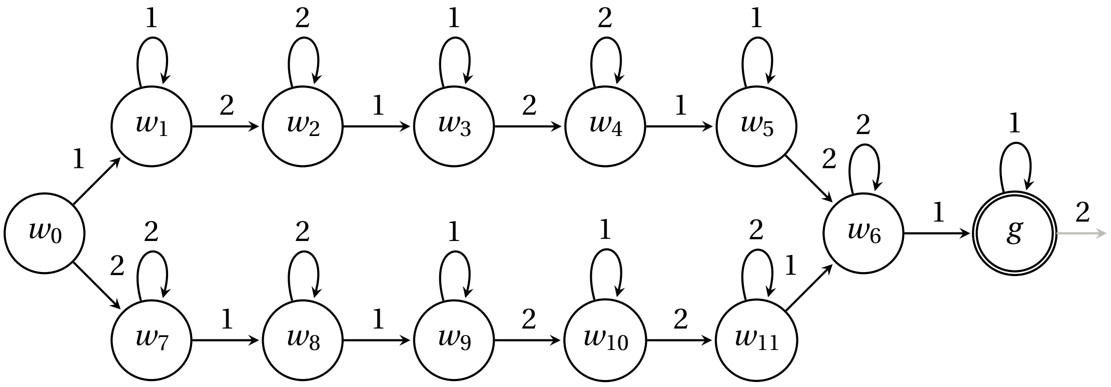


Figure 5.9: Example MDP with two different recurrent patterns. Each step has a cost of 1, except for when the agent enters the goal state, denoted g. This makes either path optimal. The top requires switching between each of both possible actions at every step, the bottom involves switching between both actions every second step.

5.3 Looking Ahead with Memory

Consider the MDP of Fig. 5.9. Again, there are two paths leading to the goal state. Both paths are optimal, and at each step along each of the paths only one action is optimal; only at w_0 does an optimal agent have a choice. Also, each action is chosen equally often along each path: three times. This means that the paths differ neither in value nor informationally, and all methods used earlier in this chapter arrive at the same policy for the agent to follow: choose uniformly in w_0, and afterwards pick the one action that is optimal.

However, if one looks more closely, there is a distinction between the paths: an agent travelling along the top alternates which action it takes at every step, whereas along the bottom it would alternate every two steps. An agent with abilities that go beyond simple reactive behaviour should be able to use this structure to alleviate its cognitive burden. In the previous chapter it was already shown that in a case like this, where the previous action predicts the next one well, some form of memory could significantly reduce the required sensory bandwidth. In this section I use the look-ahead information framework to extend on this observation, with a more concrete memory model, in the sense that it is modelled explicitly. In another sense this model is more general, as it is not restricted to retaining information across just a single time step.

5.3.1 Model

Figure 5.10 shows the perception-action loop causal Bayesian network extended with a memory. There are some similarities to the networks of Chapter 4, but the interactions are more intricate this time. Besides selecting an action, the agent also selects a memory state $m_t \in \mathcal{M}$ at each time step. In contrast to the options of the layered systems before, the influence of such a state is not limited to a single time step: the current memory state also depends on the previous, and it influences the next memory state, as well as the next action.

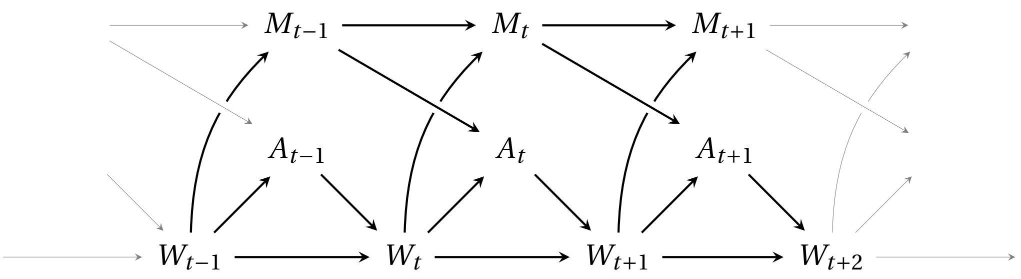


Figure 5.10: Causal Bayesian Network of perception-action loop with memory. At a time step t the agent selects an action and sets a memory state M_t, based on world information and information retained in the previously selected memory state.

Note that in this model all edges are part of the CBN; they are modelled directly. In this model, the action selection policy becomes $\pi(a_t|w_t, m_{t-1})$, the memory state-selection policy is given by the probability $p(m_t|w_t, m_{t-1})$. To ease notation, one can combine the action and memory selection: let J_t be a new variable with alphabet $\mathscr{J}$, $|\mathscr{J}| = \mathscr{M} + \mathscr{A}$, and $f : \mathscr{J} \to \mathscr{M} \times \mathscr{A}$ be a bijective function that deterministically maps each value j_t to a unique pair (m_t, a_t). Finally, let $f_m : \mathscr{J} \to \mathscr{M}$ and $f_a : \mathscr{J} \to \mathscr{A}$ return the memory state and action element of that mapping, respectively. One can now treat the combined action and memory state selection as one policy $\pi(j_t|w_t, m_{t-1})$.

The measurement for the average sensory bandwidth in this model now becomes the information that goes into selection of an action *and* a memory state, or J_t, *beyond* that which may already be available in the previous memory state, or: $I(W_t; J_t|M_{t-1})$. To find the minimal, relevant information for a given level of performance, under a memory, we arrive at the new constrained problem:

$$\min_{\pi(j_t|w_t,m_{t-1})} \quad I(W_t; J_t|M_{t-1}) \quad \text{subj. to} \quad E[U(W_t, A_t)] \leq C \tag{5.9}$$

Applying the by now familiar RI methods to this problem gives a new consistent solution:

$$\pi(j_t|w_t, m_{t-1}) = \frac{1}{\mathcal{Z}} p(j_t|m_{t-1}) \exp\left[\beta U(w_t, f_a(j_t))\right] \tag{5.10}$$

and an accompanying iterative algorithm. However, solving the problem this way does not tend to give an interesting solution. For example, for the scenario of Fig. 5.9, this method consistently gives the solution $p(m_t|\cdot) = \frac{1}{|\mathscr{J}|}$, which gives $I(W_t; M_t) = 0$ and $I(W_t; J_t|M_t) = I(W_t; A_t)$. In other words, the memory is not used at all. This can be explained by noting that the utility does not depend on the memory state, thus that $I(W_t; M_t|M_{t-1})$ is minimised unconstrained, and a trivial minimum lies at $I(W_t; M_t) = 0$.

To put it differently, this method only sees the direct, single step cost of using memory: the additional immediate demand on sensory bandwidth that it brings along. But a memory is only of use in the future; the agent should look ahead. Here I will apply the look-ahead information framework to this problem.

5.3.2 A New Definition and Solution

Similar to before, one can define a look-ahead information function, but now this time with memory. The look-ahead information of entering a state w_t, while maintaining the previous memory state m_{t-1}, is defined as:

$$\mathfrak{I}^{\pi}(w_t, m_{t-1}) = I(w_t; J_t | m_{t-1}) + \gamma \sum_{w_{t+1}, m_t} p(w_{t+1}, m_t | w_t, m_{t-1}) \mathfrak{I}^{\pi}(w_{t+1}, m_t) \tag{5.11}$$

where:

$$I(w_t; J_t | m_t) = D_{KL}\big(p(j_t | w_t, m_{t-1}) \parallel p(j_t | m_{t-1})\big) \cdot \tag{5.12}$$

$$p(w_{t+1}, m_t | w_t, m_{t-1}) = \sum_{j_t} \pi(j_t | w_t, m_{t-1}) \mathbf{1}_{j_t}(m_t) p(w_{t+1} | w_t, f_a(j_t)) \tag{5.13}$$

$$\mathbf{1}_{j_t}(m_t) = \begin{cases} 1 & \text{if } f_m(j_t) = m_t \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

Minimising this quantity over all policies, constrained to a minimum expected utility, gives us the following new self-consistent solution:

$$\pi(j_t | w_t, m_{t-1}) = \frac{1}{\mathcal{Z}} p(j_t | m_{t-1}) \cdot \exp\left[\beta U(w_t, f_a(j_t)) - \sum_{w_{t+1}} p(w_{t+1} | w_t, f_a(j_t)) \mathfrak{I}^{\pi}(w_{t+1}, f_m(j_t))\right] \tag{5.15}$$

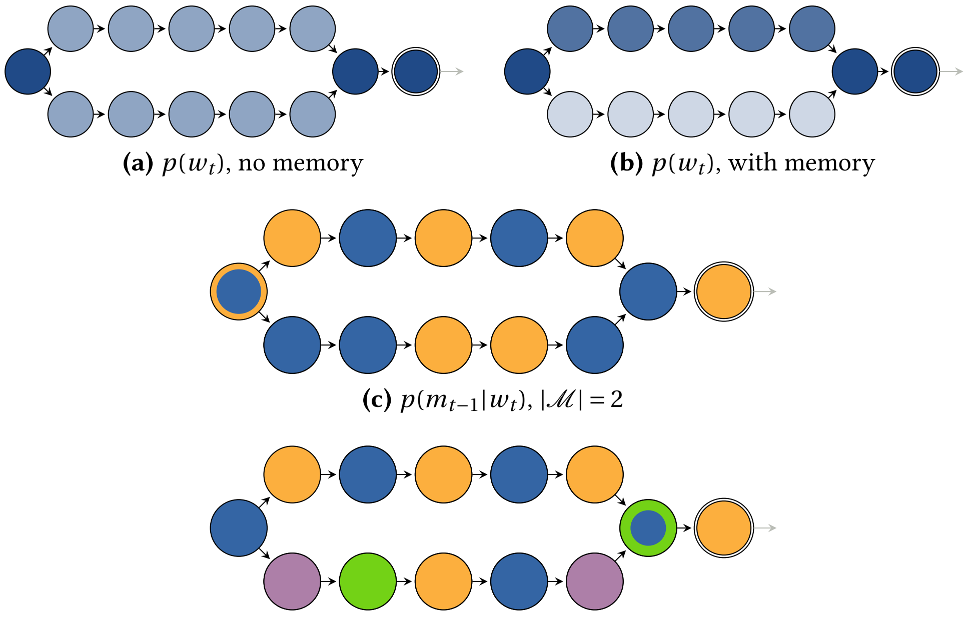
With this solution the long term effects of choosing a particular memory state are taken into account: a memory state that reduces the expected future informational burden gets a higher probability of being selected.

5.3.3 Experiment

Scenario 1: $|\mathcal{M}| = 2$

Let us now apply the new method of look-ahead information with memory to the example of Fig. 5.9. Initially, we will choose $|\mathcal{M}| = 2$, i.e. the agent can choose between two different memory states. Resulting state distributions of doing so are shown in Fig. 5.11. Firstly, for comparison, Fig. 5.11a shows the state distribution achieved with RI, C-RI, LA-RI and the earlier single step RI with memory result, in which each path is chosen with the same probability. Compare this to Fig. 5.11b, the state distribution resulting from looking ahead with memory. This time, the agent chooses the upper path just over 3 out of every 4 times.

The memory process of the agent is pictured in Fig. 5.11c. Clearly, this time the memory is extensively used: at almost every state the agent has a distinct memory state with probability 1, and this memory state differs among the world states. Only in the first state there is a mixture, which coincides with the probabilities of taking the upper route or travelling along the bottom; I will discuss this outcome later. In this outcome, each memory state is coupled to an action: blue means ‘take action 1’, and orange means ‘take action



(a) $p(w_t)$, no memory

(b) $p(w_t)$, with memory

(c) $p(m_{t-1}|w_t)$, $|\mathscr{M}| = 2$

(d) $p(m_{t-1}|w_t)$, $|\mathscr{M}| = 4$

Figure 5.11: Results of experiments in the MDP of Fig. 5.9. The top figures show the state distribution given by a policy obtained (a) without memory and (b) with memory. Whereas without memory each path is chosen 50% of the time, with memory the upper path is selected 78% of the time. Figure (c) shows the probability of either of two possible memory states being selected when the agent enters a state, Fig. (d) shows the same result for an agent with 4 possible memory states (larger radii of coloured segments denote higher probabilities).

2'. This means that the action for the current step is already defined in the previous step, which means the agent does not have to take in any information any more to choose an action.

This effect can be seen when calculating the required sensory bandwidth: $I(W_t; J_t | M_{t-1}) = 0.7$, compared to $I(W_t; A_t) = 0.875$ when no memory is used. The reason why it is not zero, is that the agent still needs some information to select a memory state: most of the time blue is followed by orange, and vice versa, but the agent must still check whether it is not somewhere where the same state is selected, i.e. in w_7, w_9, and w_{11}, and sometimes in g.

So far these results are not much different from those obtained in the previous section: there we also saw more parsimonious sensing thanks to being able, or making it easier, to predict the next action. In the next section I will show that the new look-ahead model, however, is indeed more general.

Scenario 2: $|\mathscr{M}| = 4$

One can perform the previous experiment again, but now with a memory with a higher capacity; this time, the agent can select from 4 different memory states. Firstly, the path choice in this scenario does not differ much from the previous: the agent has a similar preference for the upper path. The memory process on the other hand has become more complex, utilising the greater capacity, shown in Fig. 5.11d. Along the top path we find the same alternating choices as before, but at the bottom we see a new pattern: a repeating ordered sequence of all four memory states.

This shows that the agent manage to discover the regularity of the second path as well, which spans over more than just one time step. The memory states used here can here be interpreted as predicting the next *two* actions; for instance, purple and green mean 'do 1 now, then 1 again', and 'do 1, but then 2', respectively. Alternatively, they can be thought of as predicting both the next action *and* the next memory state. This predictive power causes a significant further drop in required sensory bandwidth: down to 0.23 bits, or just over 26% of the case without memory, and close to $\frac{1}{3}$rd of the case with half the available memory states. The fact that it still hasn't dropped to zero is explained by noting that the memory states, do not unambiguously predict the next action and memory state: the blue state has two meanings: 'take action 1 and pick yellow', and 'take action 1 and pick purple'. Some additional information is required to decide which interpretation is used.

5.4 Discussion

In this chapter I have introduced look-ahead information: the cumulative amount of future relevant information. I showed how an agent that takes into account this amount, rather than single-step relevant information, can decrease its overall bandwidth requirements, as it is aware of future informational effects of current actions. This gave rise to behaviour that caused an agent to avoid world states with uncertain outcome, to actively simplify the environment, and develop an informative memory.

These results are all grounded in the same fundamental principle of information bandwidth minimisation. They display self-organisation of behaviour driven by an intrinsic principle, rather than through externally applied pressure. Minimising informational bandwidth forces an agent to utilise the organisation of the world, and by doing so use, or create, structure in the environment that is not picked up by traditional computational decision making and learning algorithms.

Memory for instance is commonly introduced only when it is *required* to solve a problem, as in a partially observable MDP [60]: the agent needs to hold a memory of the past, making up its *believe state*, to have access to enough relevant information to select the best action. In this chapter however the world is fully observable, but the parsimony pressure *still* gives rise to memory that predicts the world state well. Thus, this approach shows how memory can be bootstrapped without presupposing a requirement, possibly opening up solutions to new problems where memory *is* required. Similarly, in the obstacle removal scenario, the observed behaviour is not necessary performance wise, but may open up tasks where this already acquired behaviour *is* beneficial. This effect, where information minimisation gives rise to beneficial properties, will come up again, and be discussed in more detail, in Chaps. 7 and 8.

The effects observed in this chapter are not limited to the scenarios that were used; whether they appear only depends on whether there are local decisions that the agent can make that affect the state distribution enough to make the local policy on average more similar to the global average action choice. The focus is on 'enough': if for instance the world of Fig. 5.3 would have been reduced, the chance of jumping into the water from a lily pad would have been lower, reducing the effect of that happening on the average information intake. If it is reduced below the effects of deviating from the average policy when trying to ensure a land route is taken, and the preference for specific routes disappears as with using single step RI. Similarly for the obstacle removal scenario, if the obstacles would come back more frequently, the future effect of the immediately informationally costly action of removal may become too low to be worth it. This same result is expected when reducing the discount factor $gamma$: if it is set low, and thus the agent values immediate information reduction more, the effect of looking ahead is reduced.

Some authors have arrived at results related to those presented here. Firstly, look-ahead information is highly related to *information to-go* [105]. This measurement also gives a cumulative future information, given as a Bellman-type recursive equation. However, the information to-go includes the environmental response, determined by the transition model, *as well* as the agent's decision complexity. The look-ahead information is derived by taking only the latter term, *plus* the direct informational cost of the current action a_t. This addition is important, since as the examples of this thesis show that reduction of future informational cost often requires a higher immediate cost.

A similar result as in the memory experiments is arrived at in an informational treatment of interactive learning [94], in the form of an internal model that predicts the future state of the world. In that work however, this is posed as an a-priori goal, by explicitly maximising the information between the model built by the agent and the world state. In this chapter, the predictive memory arises on the other hand from a drive for information *minimisation*. Furthermore, the look-ahead information framework would reward memory that is relevant to any future time step (if it is maintained in memory for that long), whereas the referenced methods are limited to predicting a single time step. In any case comparison is difficult as the authors do not provide empirical experimental results.

The parenthesised note in the previous paragraph hints at some questions that are still open. Mainly, I have focused solely on the cost of sensory bandwidth, implying that memory, and its input and output channels, are free. In the next chapter I will discuss a trade-off between different informational costs, an interesting endeavour may be to perform a similar analysis by introducing costs on memory usage. A difficult question here however is how to select a good cost function, specifically on the cost of maintaining static memory over time, versus a dynamically changing memory.

Chapter 6

Relevant Goal Information

6.1 Introduction

Up until now we have only treated a simplistic model of behaviour, where the current state of the environment supplies all information needed to drive an agent's actions. Although this simple model can fit a wide range of behaviour, it does not fully capture more complex behaviour repertoires, where the current state, or even the full state history, of the environment is not enough to decide upon an action. Finding myself entering the kitchen does not give enough information to know whether I should make tea or a sandwich. The decision on which behaviour is applicable here depends on the state of an internal drive, e.g. on the internal states of being hungry or thirsty.

There are many drive states that one could identify. What they have in common is that the behaviour that is performed under their influence can be said to be performed in order to achieve something specific. As Kupfermann et al. [55] put it:

> Drive states [...] serve three functions: they direct behaviour toward or away from a specific goal; they organise individual behaviours into a coherent, goal-oriented sequence; and they increase general alertness, energising the individual to act.

They go on to describe a range of drive states that can be modelled with servo-control processes, such as hunger, temperature, and some forms of addiction. In these processes, one could ask whether the 'specific goal' that is mentioned is a real concept, identifiable in the agent, or rather a high level human modelling tool. Neural correlates of goal representations *have* been identified in human brains, both in the context of generating behaviour [69], and of observing goal-directed actions [38]. Similar results have been found for monkeys [82]. In any case, I would argue that any organism that is able to switch between alternative coherent behaviour patterns, beyond reactive responses to external stimuli, must maintain some internal representation that can be related to a goal to regulate these patterns. This argument goes at least as far down as to the complexity scale of the nematode worm *C. elegans* [58].

In this chapter I will extend the framework introduced so far to incorporate goal-directed behaviour. In line with the previous chapters, I am interested in the necessary properties of the agent's internal cognitive facilities that generate goal-directed behaviour. Again, the underlying proposal is that these facilities are not unbounded; the extension of the model introduce several new possible constraints in different sections of the cognitive processes. In this chapter, I propose and formulate several of such constraints, derive what

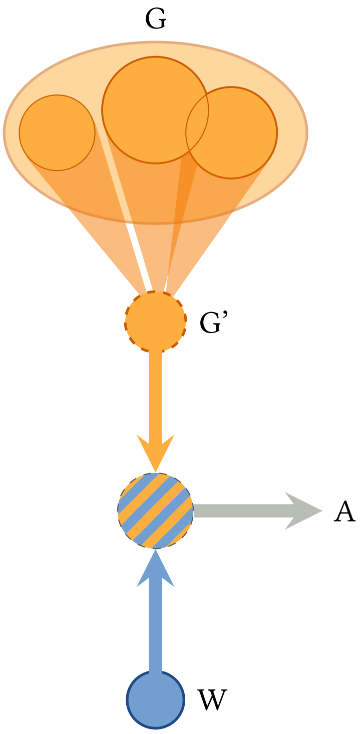


Figure 6.1: Overview of cognitive model of goal-directed behaviour underpinning this work. The agent has full, persistent knowledge about the current goal, represented by G. As denoted with the smaller disks, different aspects of this knowledge is used at different times while acting to achieve the goal. The currently relevant part is loaded in a dynamic working memory, G', which has a limited capacity. There, this goal knowledge is combined with sensory information about the world state W, and the agent chooses its actions A based on this combination.

the tightest constraints are that still allow an agent to perform well, study how they influence goal-directed behaviour, and, conversely, investigate to what extent they drive the structure of goals and the environment to be reflected in the agent's cognitive processes.

An informal sketch of the cognitive model of goal-directed behaviour that I will use is given in Fig. 6.1. It shows that State information is now combined with information about the internal motivations of the agent, or *goal information*, to generate actions. From this model, we can develop an extended formal description of the perception-action loop, with which we can again pose cognitive constraints as informational limitations.

In the remainder of this chapter we will study the properties of goal-directed behaviour under several such constraints. Firstly, we start with some observations resulting from such limitations: a limited working memory leads to ritualised behaviour patterns, and, to some extent, sensory and goal information can be substituted for each other if they have different costs. Next, we take a look at the dynamics of information that is loaded and unloaded in the constrained working memory described above. We find that these dynamics reflect salient and natural transition points in the environment. Finally, we show how restricting the bandwidth on a persistent memory channel, combined with a notion of effectiveness of use of that channel, allows one to identify the 'dual' of the above transitions, namely helping the agent to decompose the world into a natural abstraction.

When we talk about goal-directed behaviour, we imply the ability to make a distinction between different goals: the agent selects one goal and then generates behaviour by selecting actions that should achieve this goal. This behaviour is of course distinct for differing goals. The following section develops a formal model in which this idea is embedded.

6.1.1 Base Model: a Family of MDPs

The formal model of goal-directed behaviour that I will employ is that of a family of MDPs, which I will introduce here. This model is similar to how *Multi-Task Learning* problems are defined by Taylor et al[101].

In this family, each MDP consist again of state and action sets, state transition probabilities, and a reward function, as described in Sec. 2.3.1. The first three of these are equal and fixed throughout all MDPs; the distinction between them is expressed in the difference in which actions are preferred in certain states, which is given by the reward function. If we denote a specific MDP with an index g, out of the set of indexes $\mathscr{G}$, the reward function belonging to this MDP is then given as $R^a_{gww'}$. Each MDP is then defined by the tuple $< \mathscr{W}, \mathscr{A}, P^a_{ww'}, R^a_{gww'} >$.

When presented with a set of MDPs, the problem is now to find a policy that maximises the expected reward over the full set. To be able to do so, an agent can no longer keep a single policy for each state. Instead, it will have to select actions based on both the current state *and* the currently selected MDP. In other words, the agent's decision making process is now described by a probability distribution $\pi(a|w, g) = Pr(A_t = a|W_t = w, G = g)$.

For each MDP and an according policy, a value function $V^\pi_g(w)$ and a utility function $U^\pi_g(w, a)$ as before (c.q. Sec. 2.3.1). The generally expected value than is equal to the sum $\sum_{g\in\mathscr{G}} p(g)V^\pi_g(w)$, where $p(g)$ is the probability of MDP being currently selected. In the remainder the distribution over MDPs within a set is set to be uniform.

6.1.2 Specific Example, and Generality of Concepts

As an example of a goal-directed scenario that can be described as a family of MDPs, we use a grid-world navigation problem, which is a popular test scenario in RL literature [97]. The specific scenario is the same as one used by ŞimŞek and Barto [88], shown in Fig. 6.2, where an agent has to navigate across a number of rooms to reach some goal location. The set of world states $\mathscr{W}$ consists of the possible locations of the agent, i.e. corresponding to the empty cells. The action set $\mathscr{A}$ contains four possibilities: move north, east, south, or west. A specific goal consists of navigating to a certain cell, and the goal set $\mathscr{G}$ contains all possible cells as goal state. For each goal, a reward function $R^a_{gww'}$ is constructed as follows: each transition gives a reward of -1, except when entering the goal cell, where the reward is 0.

This means that in this scenario each goal corresponds to a target world state that should be reached. I focus on such scenarios because goal state-directed navigation is a fundamental and often studied problem, the results of which are easy to interpret and

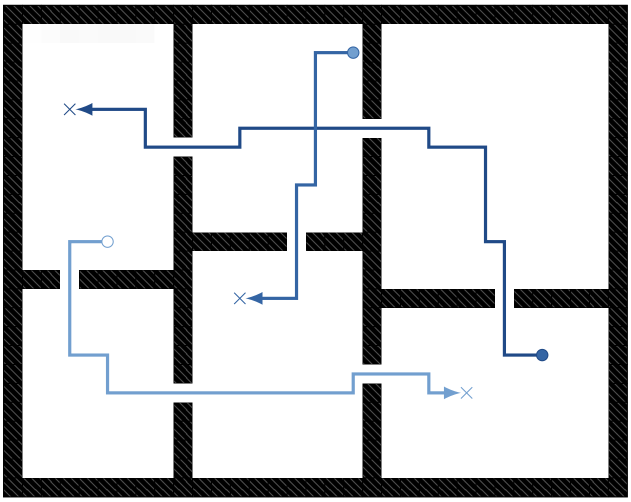


Figure 6.2: Example goal-directed scenario: Grid-world navigation. The world state consists of the location of the agent (circles), and here a goal consists of reaching a certain location (crosses). This figure shows three possible goal-directed action sequences, with possible start states and (optimal) solution paths.

relate to the work of the previous sections. However, note that in general, scenarios do not necessarily have a direct correspondence between goal and a desired world state. As quoted before from Kupfermann et al. [55], a drive state could actually direct behaviour *away* from some state. Or, in some scenarios a relative change of state may be what determines desirable actions, such as obtaining more food in a foraging task. Therefore, it is important to keep in mind that a goal g is merely an index on a set of (possibly random) reward functions, and that the fundamental principles of our approach apply unchanged to this more general concept of goal-directedness.

6.1.3 Information

The formulation of a family of MDPs can be captured graphically as a graph of its elements and how these elements influence each other, the persistent overall goal is represented by a random variable G, which is the same throughout time. The variables S_t and A_t, on the other hand, correspond to the fast-changing states of the sensors and the actuators of the agent during its run. Connecting these variables based on their causal interactions results in the network of Fig. 2.5a, which models the perception-action loop.

As the edges of the network show, information about the world state is no longer sufficient to select an action, in contrast to the models of the previous chapters: information about both the current state *and* the goal is required. One can again as before apply information theoretical measurements to this new network. For the investigation for this thesis, it is for instance of interest to know the amount of information from the combination of world state and goal as information sources that is actually required on average to select an action. This can be determined with the mutual information between the combination of state and goal variables, and the output action: $I(W_t, G; A_t) = \sum_{w_t,g,a_t} p(w_t, g, a_t) \log \frac{\pi(a_t|w_t,g)}{p(a_t)}$. Finally, we can determine the informa-

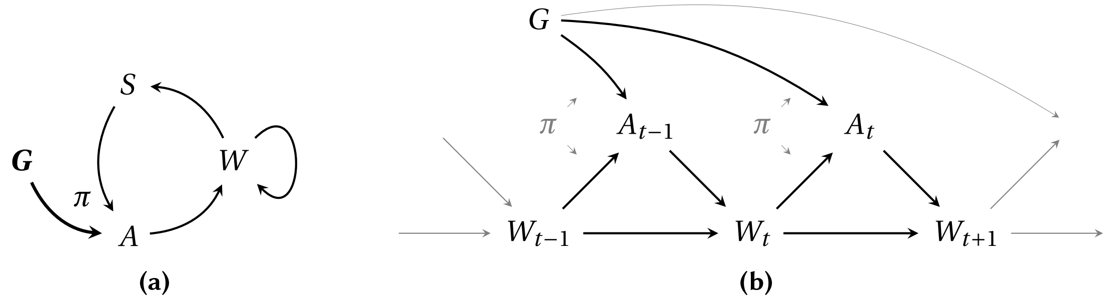


Figure 6.3: (a) Perception-action loop (arrows in (a) are to be understood informally as scheme of influence between the different variables in the system, while the arrows in (b) are to be understood in the formal framework of Bayesian Networks.) (b) Causal Bayesian network of the perception-action loop unrolled in time. As before, it is assumed that the agent in theory can fully observe the world, such that the world and sensor state at time t can be contracted into the variable W_t. The action selected by the agent, according to its policy π, by A_t. The current goal is determined by G, which persists throughout time.

tion that is directly given specifically by one of these inputs. For instance, the conditional mutual information $I(G; A_t|W_t)$ gives how much information about the goal is reflected in the chosen action, given knowledge about the current state. The conditioning on the current state is important, since there is a strong interplay between state and goal information: knowing the goal is often not very informative if the current state is not known. This means concretely that we can generally observe the relation $I(G; A_t) < I(G; A_t|W_t)$.

6.2 Informational Bounds in Goal-Directed Behaviour

The framework of the previous section allows us to make the concepts and goals of this chapter more concrete. In line with the previous chapters, here I am interested in the effect and properties of cognitive constraints on the behaviour of agents, and the informational framework offers a natural formulation of such constraints as limited bandwidths of the channels between information sources and other components of the behaviour generating system.

With the current goal identified as a required information source, one can find how much goal information *needs* to be accessed to reach a certain level of performance. Applying the method of relevant information, this again is formulated as a minimisation problem of the following form:

$$\min_{\pi(a_t|w_t,g)} I(G; A_t|W_t) \quad \text{subj. to:} \quad E[U^{\pi}_{G}(W_t, A_t)] \geq C_U, \tag{6.1}$$

This problem gives a new self-consistent solution, that can be used in the fixed point iteration, Blahut-type algorithm to find the minimum[1]. In parallel to the relevant state information studied before, the solution of this problem is dubbed the *relevant goal information (RGI)*, as it is the amount of information about the current goal that on average is

[1] see 6.6.2 for derivations of the solution to this problem, as well as to other problems introduced below.

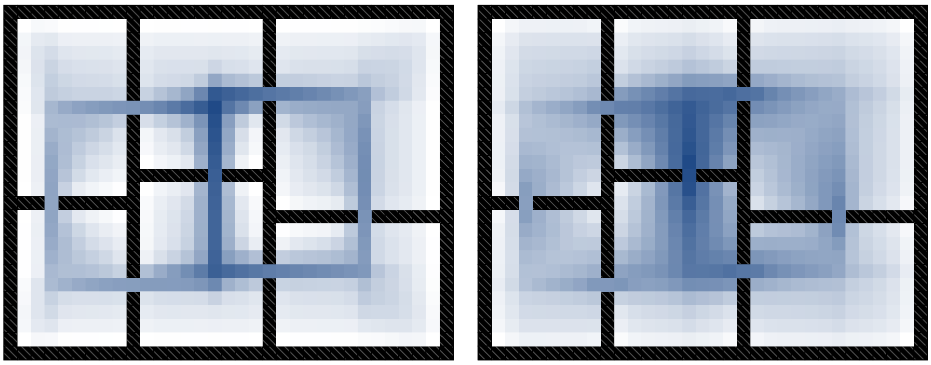

(a) RGI **(b)** VI

Figure 6.4: State distribution resulting from policies obtained by (a) Relevant Goal Information minimisation, and (b) traditional Value Iteration. The shading of a cell is proportional to the log-likelihood $\log(p(w_t))$ (chosen to easier discern the difference between low probabilities) of the agent occupying that cell at a random point in time. See Sec. A.2.1 for details on how the state distribution is determined.

relevant to achieving a certain level of performance; any other available information beyond this amount is irrelevant and does not need to be accessed when selecting an action. In other words, a working memory needs to maintain on average only this amount of information, and as such it gives a necessary lower bound on the goal-information capacity of such a working memory.

In general, in a scenario where a small set of actions is available to achieve a large set of varying goals, such a memory can be significantly smaller than the persistent memory that maintains all goal information. This can be seen by realising that the relevant goal information is bounded by the maximum action entropy: $I(G; A_t|W_t) \leq \log(|\mathscr{A}_t|)$, which in general is significantly less than the goal entropy $H(G)$, except for simple cases where only a few goals have a high probability of becoming active. In our navigation scenario we can for example determine the maximum average goal-information capacity that the agent's working memory ever needs, by performing the relevant goal information optimisation in the special case of $C_U = \max_\pi E[U_G^\pi(W_t, A_t)]$, i.e. when we quantify the required goal information for achieving the best possible reward.

Doing so, we find a policy requiring on average 1.06 bits of goal information, which is indeed significantly less than the total of $H(G) = 9.4$ bits that are in principle required to identify the complete goal. So, only about 1 bit of information about the goal needs to be available on average in the agent's working memory to select an action; the agent does not have to process the full 9.4 bits at once.

6.2.1 Observation 1: Ritualised Behaviour

Now that we have established that the agent can get away with a very small working memory, we study what effects such a small working memory has on the organisation of the agent's behaviour. One way is to determine the marginal state distribution $p(w_t)$ resulting from this policy. This distribution gives the possibility of finding the agent at any given cell at a random point of time, and as such provides an indication of the paths an agent tends to take towards its goal.

When we determine this distribution in our example pictured in Fig. 6.4a[2], given the minimal optimal policy found in the previous section, a lattice pattern of 'preferred states' is visible: there is a relatively high probability of finding the agent in a direct line with one or more of the doorways. For comparison, Fig. 6.4b shows the state distribution induced by the policy that uniformly chooses between optimal actions, without also taking informational optimality into account

This preferred paths can be explained as a direct result of the posed informational constraints, and thus expected generally in other scenarios, by noting that the policy minimises the difference between the a-priori action entropy and the entropy given the goal: $I(G; A_t|W_t) = H(A_t|W_t) - H(A_t|W_t, G)$. This minimisation is achieved when the local policy at different states for different goals is similar to the policy averaged over all goals, i.e. when $p(a_t|w_t, g) \approx p(a_t|w_t)$. In other words, optimisation aims to render the behaviour for different goals as similar as possible, which leads to a form of 'ritualised behaviour' that moves the agent along the preferred trajectories that are visible in Fig. 6.4a.

6.2.2 Observation 2: Independent Types of information can be Interchanged

As Fig. 6.3 shows, the agent's behaviour is determined by both the goal, *and* state information input. We can construct a minimisation problem similar to that of (6.1) to find the *relevant state information*, or the lower bound on the working memory capacity for state information. Moreover, we can find the *combined* trade-off between state and goal information and performance, formulated as the following problem:

$$\min_{\pi} \alpha I(W_t; A_t|G) + (1-\alpha) I(G; A_t|W_t) \quad \text{subj. to:} \quad E[U_G^{\pi}(W_t, A_t)] \geq C_U. \tag{6.2}$$

Here, $\alpha \in [0-1]$ allows us to weigh the two types of information. For instance, in the case that the capacity is limited due to some cost on information processing, e.g. due to energy constraints, this parameter lets us model the relative size of such costs.

From our earlier discussion of the relevant information as *the* minimum amount that is required to achieve a certain level of performance, one can expect that the relative cost does not influence the result: this amount is necessary, regardless of how costly it is. However, it turns out that this only holds for optimal behaviour; for sub-optimal policies the more expensive type of information may be traded off against the other type.

To illustrate this, Fig 6.5 shows for instance the results of solving (6.2) for the full possible ranges of α and C_U, in a grid-world navigation scenario using the simpler 6-room world of Fig. 6.9 in order to keep finding the large number of trade-offs tractable. As denoted with the solid lines, different trade-offs of the two types of information are possible in order to achieve a given level of performance, depending on their relative cost. In other words, knowing the state of the current world in higher precision allows you to consider your goal in less detail. Alternatively, if obtaining or processing sensory information is

[2] See A.2.1 for details on how this distribution is determined

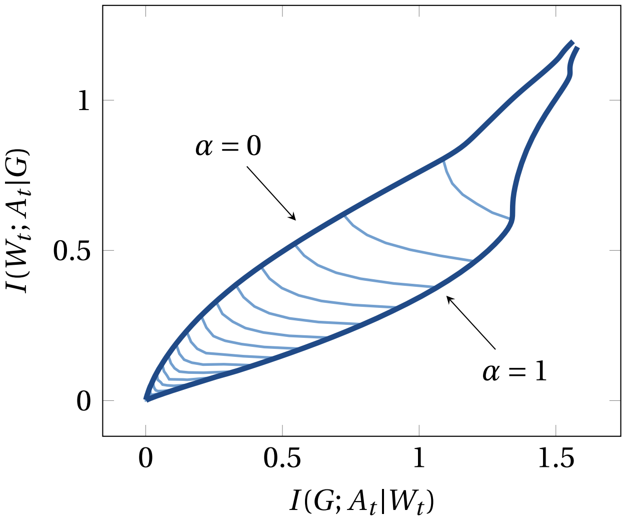


Figure 6.5: Trade-offs between state information, goal information, and utility, obtained for a 6-room navigation scenario. The thin isolines show trade-offs between sensory and goal information that achieve the same level of performance. The thick outer lines mark the boundaries of the sensory-motor optimised area, in which all optimal state-/goal-information/performance trade-offs lie.

more costly or has stronger constraints, the agent *must* employ a larger goal information bandwidth.

Finally, note that the area within the dashed curves encompasses the full range of *sensory-motor optimised behaviour*: any policy outside of this set is sub-optimal for *any* combination if bandwidth constraints and desired performance, as there is always another policy that achieves better performance with the same amount of information, and/or decreases the information bandwidth requirements on one or both of the sources of information without deteriorating performance.

6.3 Working-Memory Dynamics

In the previous part we have looked at how much the capacity of a working memory in a goal-directed agent can be constrained, by deriving the lower bound on the amount of information that must be retained to behave successfully. This amount is an *average* over all world states. This amount can be broken down for each state, which shows that the actual relevant information changes over time. In Fig. 6.6 this breakdown is shown, as it visualises the quantity $I(G; A_t|w_t)$ for each state w_t. Some patterns are visible in this figure, most notably that the amount of information on the outside of the world is significantly lower than around the centre. This can be explained by the fact that most of the time, when the agent is in these outer states, it is on the way to a goal state in this outer region. This means that most of the time the agent will not choose a sub-optimal action that would backtrack and bring it back inwards again. Knowing this, the a-priori action entropy is lowered and the relevant goal information is limited, compared to the inner states where there is no (or at least much less) such a-priori knowledge of the action

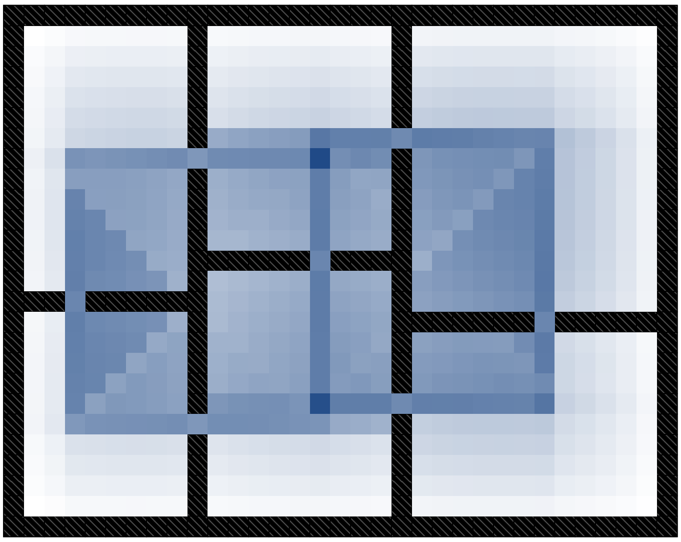


Figure 6.6: Relevant goal information per state. The darkness of the shading of each cell is proportional to the value of $I(G;A_t|w_t)$, where w_t denotes the state of the agent occupying that cell.

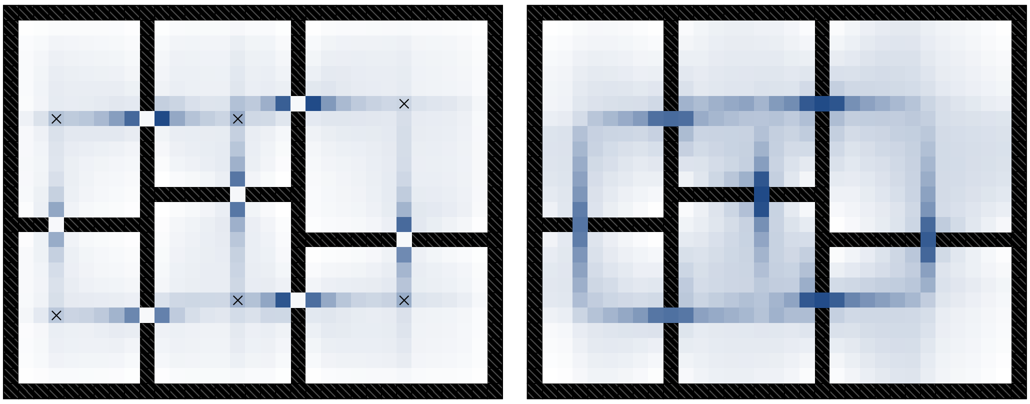
distribution.

Beyond this distinction between inner and outer states, one can make the more general statement that in some states more information is required than in others. Furthermore, even when the amount is the same, the actual content can still be different. This means that the agent must continuously 'download' different parts from the persistent complete goal memory into its working memory, and can 'unload' parts that are no longer needed. As one would expect from the name, the working memory is dynamic, and in this part we will study its dynamics.

Again we will adopt the viewpoint of constrained cognition, and find the lower bounds on the information in and out flow of the working memory. We can expect that these are lower than the total amount of relevant information: it is likely that part of the goal information that was relevant at one time step is still relevant the next, and thus can be retained in the working memory. This means that the minimum amount of *new* information that must be transferred from the static memory at a certain state, is that which is relevant to the current action selection, *beyond* that which was relevant to past decisions and therefore possibly already is available in the working memory. These past decisions are reflected by the state-action pairs encountered thus far, which we will denote as the random variable $\mathbf{E}_{t-1} = (W_0, A_0, W_1, A_1, \ldots, W_{t-1}, A_{t-1})$. The amount of new goal information transfer needed at a state w_t, which we will denote with $I_G^{new}(w_t)$, is then determined by:

$$I_G^{new}(w_t) = I(G;A_t|\mathbf{E}_{t-1}, w_t) \quad = \sum_{g,a_t,\mathbf{e}_{t-1}} p(g, a_t|\mathbf{e}_{t-1}, w_t) \log \frac{p(a_t|g, \mathbf{e}_{t-1}, w_t)}{p(a_t|\mathbf{e}_{t-1}, w_t)} \tag{6.3}$$

In a similar way, one can define the dual quantity of $I_G^{old}(w_t) = I(G;\mathbf{E}_{t-1}|A_t, w_t)$. This quantity gives the amount of *old* information that was relevant and available (at some



(a) $p(w_t)I_G^{new}(w_t)$ **(b)** $p(w_t)I_G^{old}(w_t)$

Figure 6.7: Goal information transitions in a 6-room grid world navigation task. Figure (a) shows the amount of *new* goal information needed to select an action for each state, $I_G^{new}(w_t) = I(G; A_t|\mathbf{E}_{t-1}, w_t)$. Figure (b) shows the amount of *old* goal information no longer needed to select an action for each state, $I_G^{old}(w_t) = I(G; \mathbf{E}_{t-1}|A_t, w_t)$. Values are weighed with $p(w_t)$, to give the contribution at each state to the total average information transfer. Darker shades denote higher values. States that lie on the crossing points between two doorways, which show local maxima in the amount of required new information required, are marked with × in (a).

point) in the working memory in the past, which is no longer relevant and can thus be offloaded to make space in a constrained working memory

6.3.1 Observation 3: Two Types of Natural Sub-Goals

In Fig. 6.7 we show the results of determining the two goal information-transfer quantities introduced above. They are shown weighed by the state distribution $p(w_t)$, to give the contribution at each state to the total average information transfer over all states. In both figures a salient pattern appears: around the doorways, both the amount of new information that needs to be loaded into the working memory ($I_G^{new}(w_t)$, Fig. 6.7a), and the amount that is no longer relevant and thus can be unloaded ($I_G^{old}(w_t)$, Fig. 6.7b) is high.

These results can be understood by considering what goal information an agent needs to be able to select an action. Imagine an agent navigating from one room to a goal in the next room. Before transitioning into this room, the only goal knowledge that the agent needs to use is that the goal is not in the current room, and which doorway to go through to get into the next room on its way to the goal. Once in this doorway, this information can be discarded. Next, when entering the following room, the agent now needs to load and consider the new information of where in the room the goal actually is.

More concretely, we can expect to find this kind of transitions at states where the local policy is independent of past experiences, which is especially the case at states that connect separate strongly connected areas. As we have discussed previously [109], these states tend to coincide with the top-down definition of sub-goals as commonly used in

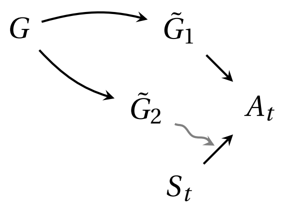


Figure 6.8: Causal Bayesian Network with goal-information bottlenecks $\tilde{G}_1$ and $\tilde{G}_2$. The grey wavy arrow from $\tilde{G}_2$ denotes that the information in this bottleneck does not directly influence a specific other variable, as is represented by the normal edges, but rather modulates the state information pathway to enable the selection of the relevant state information.

the field of reinforcement learning [88]. The results of this section indicate that such concepts can expressed immediately in terms of cognitive (informational) requirements of goal-directed behaviour, and arise naturally when one considers how the task needs to be organised in the case of tight cognitive resources.

However, there is an additional benefit of grounding these kind of concepts in internal processing requirements, rather than external properties of the world. Consider the comparison of $I_G^{new}(w_t)$ with the novel results for $I_G^{old}(w_t)$ in more detail. Local maxima of $I_G^{new}(w_t)$ can be found not only at obvious locations, such as doorways, but also at states at the crossing between different doorways, marked with × in Fig. 6.7a. Interestingly, there is no corresponding peak at these states in the amount of information that can be discarded. So, at these states the transition consists mostly of requiring additional goal information to that already available in the working memory. These then constitute a second distinct type of salient transition points, arising from our framework, in addition to those traditionally regarded as sub-goals, and these two types can be intrinsically distinguished via their goal information signature.

6.4 Explicit Constraints on Goal-Information Pathways

Up to this point, we considered informational limits of cognition implicitly, by determining the lower bounds on the amount of information that a working memory must contain, and on the amount that must be loaded and could be unloaded from this memory. We will now study the scenario where explicit constraints are placed on the bandwidths of goal-informational pathways.

The first pathway that we consider is the one which transfers the goal information that directly influences the selection of actions. This is the aforementioned relevant goal information, measured with $I(G; A_t|W_t)$. A restriction on the capacity of this pathway is modelled by placing a random variable $\tilde{G}_1$ between G and A_t, as shown in Fig. 6.8. The bandwidth constraint is then established by a strongly reduced cardinality of the alphabet of $\tilde{G}_1$, i.e. $|\tilde{\mathcal{G}}_1| \ll |\mathcal{G}|$.

The *Information Bottleneck (IB)* method [104] covers exactly this problem: the variable $\tilde{G}_1$ constitutes a 'bottleneck' through which the relevant information must be squeezed. In this paradigm, the problem of transferring as much relevant information as possible through this bottleneck, while limiting the amount of irrelevant information that is captured in it, is posed as another constrained information minimisation:

$$\min_{p(\tilde{g}_1|g)} I(G;\tilde{G}_1) \quad \text{subj. to} \quad I(\tilde{G}_1;A_t|W_t) \geq C_{I_1} \tag{6.4}$$

This problem has a form similar to the relevant information problem (6.1). However, here the relevance of the information is determined purely informationally, by the lower bound imposed on the retained information. The solution of (6.4) results in a distribution $p(\tilde{g}_1|g)$ that induces a stochastic mapping, or soft clustering of the set of goals into a smaller set, and as such can be regarded as a coarser abstraction that captures the most relevant structure of a goal description.

The second pathway for goal information that we examine is more subtle. As we have mentioned previously, goal and state information display an intricate interplay; one can not be considered separate from the other, as one modulates the other type of information to 'unlock' the relevant aspects. As noted earlier, the effect of goal information on the relevance of state information is quantified by the increase of $I(W_t;A_t|G)$ with respect to $I(W_t;A_t)$. Our second pathway can be seen as one which transfers the goal information that is needed to achieve the necessary modulation of the state information channel.

The bottleneck variable $\tilde{G}_2$ shown in Fig. 6.8 models an explicit bandwidth constraint on this second pathway. Whereas $\tilde{G}_1$ must capture information relevant to A_t, the success of which is measured with $I(\tilde{G}_1;A_t|W_t)$, $\tilde{G}_2$ must capture information that increases the relevance of W_t, measured with $I(W_t;A_t|\tilde{G}_2)$. From this comparison we construct the following IB-type problem to find the optimal bottleneck distribution for this pathway:

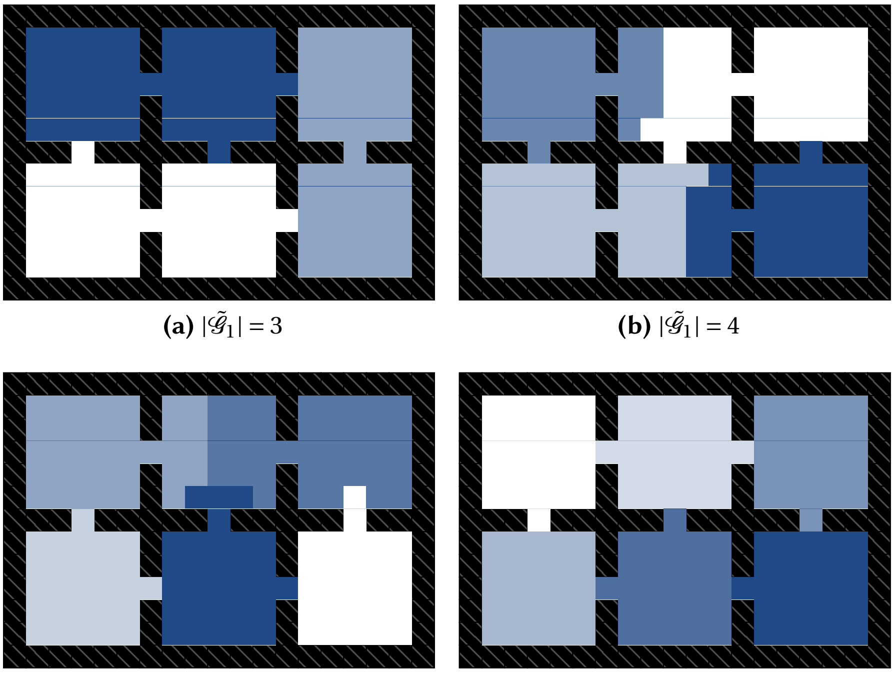
$$\min_{p(\tilde{g}_2|g)} I(G;\tilde{G}_2) \quad \text{subj. to} \quad I(W_t;A_t|\tilde{G}_2) \geq C_{I_2} \tag{6.5}$$

6.4.1 Observation 4: Natural Abstraction

Firstly we will study the primary pathway, constrained with bottleneck $\tilde{G}_1$, and solve (6.4) to find the goal mapping induced by this bottleneck on the pathway. Figure 6.9 shows such mappings found for different capacities of the bottleneck variable in a 6-room grid navigation scenario with the lower bound C_{I_1} fixed as high as possible (see 6.6.2 for more details), such that the clustering becomes most informative.

One result to note is that the stringent lower bound results in a hard clustering: each goal state is deterministically mapped to a single element in $\tilde{\mathcal{G}}_1$. Secondly, the mapping adheres to the local connectivity of goals: goal states in the same cluster are connected directly in the transition graph of the MDP.

Moreover, the clustering also attempts to adhere to the physical boundaries of the environment: goal states within a single room tend to be clustered together, with transitions between clusters tending to be at the doorways. Especially when the number of clusters coincides with the number of rooms, i.e. for $|\tilde{\mathcal{G}}_1| = 6$, the decomposition of the world induced by the mapping seems to match the environment especially well. We can quantitatively identify a number of clusters as the most adequate, by introducing a measurement for the degree of effective use of the restricted bottleneck capacity.



(a) $|\tilde{\mathscr{G}}_1| = 3$

(b) $|\tilde{\mathscr{G}}_1| = 4$

(c) $|\tilde{\mathscr{G}}_1| = 5$

(d) $|\tilde{\mathscr{G}}_1| = 6$

Figure 6.9: Goal clusters induced by the bottleneck $\tilde{G}_1$ on the primary goal-information pathway in a 6-room grid world navigation task. Figures (a) to (d) show the mappings for increasing cardinality of the bottleneck variable.

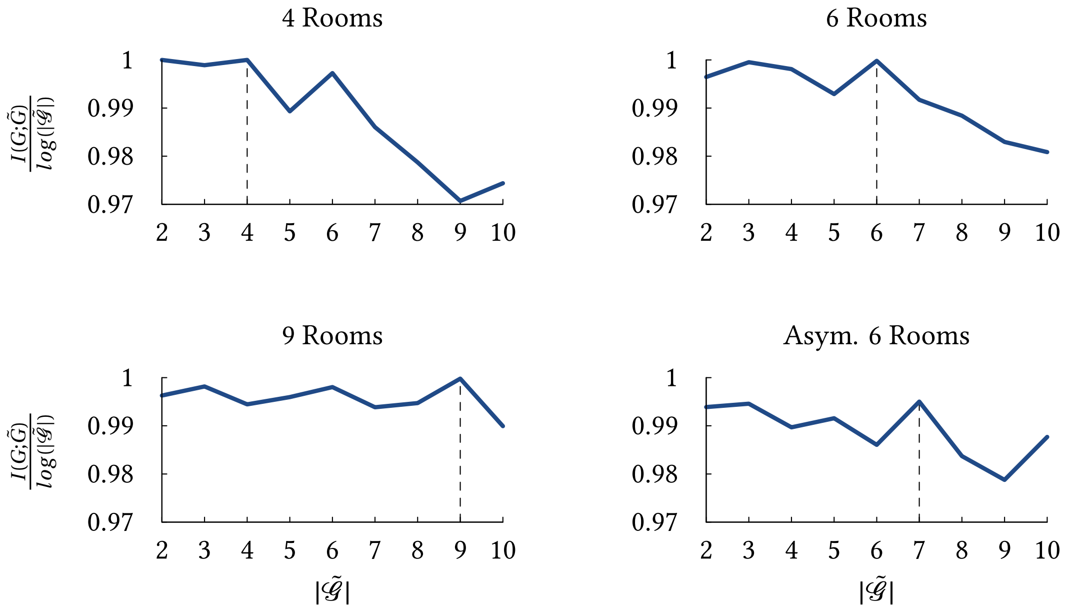


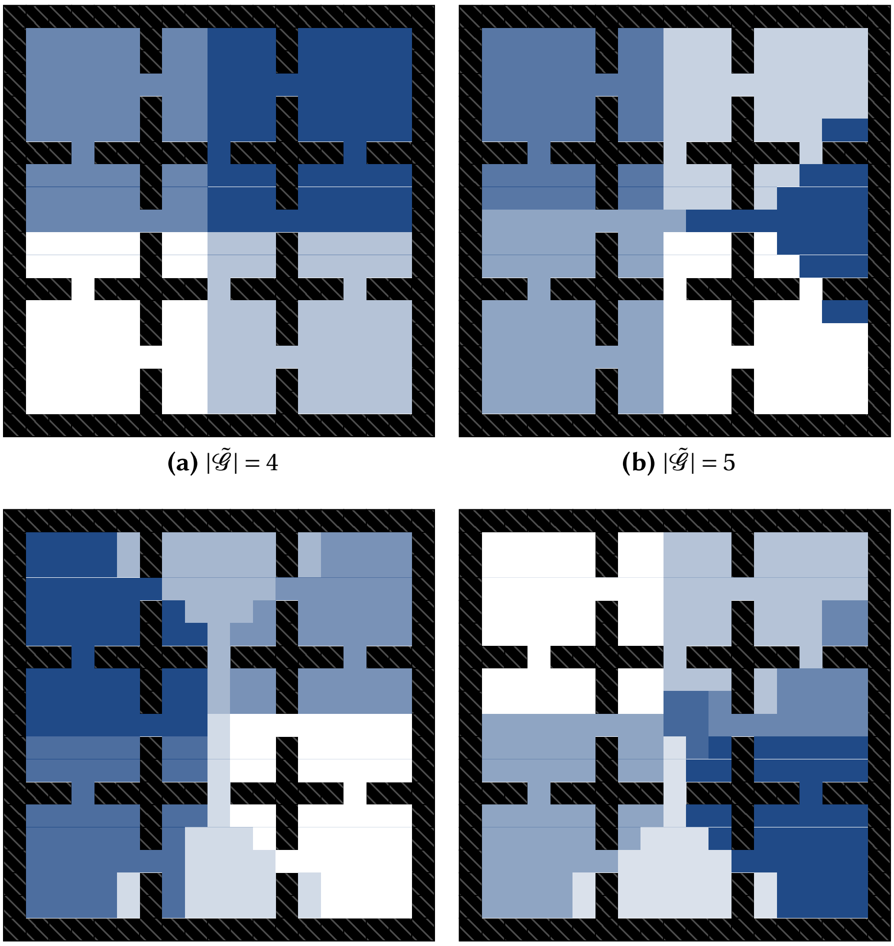
Figure 6.10: Effectiveness of clustering, measured by the ratio of the amount of goal information captured in the clustering, $I(G;\tilde{G}_1)$, to the total capacity of the bottleneck, $|\tilde{\mathscr{G}}_1|$, plotted for different cardinalities of the bottleneck variable. Values closer to 1 indicate better use of the available capacity. The maximum for each scenario is marked with a dotted line.

The maximum capacity of the bottleneck is given by $\log|\tilde{\mathscr{G}}_1|$; e.g. if the cardinality of the variable is 4, these four clusters could maximally capture 2 bits of goal information. The bandwidth *use* is given by the actual amount of goal information that is reflected in the clustering, $I(G;\tilde{G}_1)$. The ratio of these quantities, $0 \leq \frac{I(G;\tilde{G}_1)}{\log|\tilde{\mathscr{G}}_1|} \leq 1$, then gives a quantitative indication of the effective use of the available capacity. The residual quantity $1 - \frac{I(G;\tilde{G}_1)}{\log|\tilde{\mathscr{G}}_1|}$ is known as the *Shannon redundancy*, as it gives the ratio of the bandwidth that is not used and thus is redundant; see [11] for a discussion of its importance in neuro-science.

In Fig. 6.10 we show the result of determining the above ratio ratio for several room-navigation scenarios, for a range of bottleneck sizes. It is clear that some sizes result in more effective use of the constrained pathway than others, and that a less stringent constraint does not necessarily result in less redundancy. As a matter of fact, we find that for the 4, 6, and 9 room scenarios the maximum effectiveness is achieved when the number of clusters exactly matches the number of rooms. For instance, in the case of the 6 room scenario, the clustering that achieves maximum effectiveness is indeed that of Fig. 6.9d.

We suggest that these maximally effective factorizations, induced by an explicitly constrained information pathway, supply *natural abstractions* of the environment that capture the most relevant structure that is available most effectively.

The graphs also show other interesting factorizations as secondary and tertiary peaks. Here, the bottleneck, though not able to assign a single cluster to each room, still adheres mostly to room boundaries, such as for $|\tilde{\mathscr{G}}| = 3$ in the 6-room world (cf. Fig 6.9a), and $|\tilde{\mathscr{G}}| = 3$ and $|\tilde{\mathscr{G}}| = 6$ in the 9-room world. Interestingly, for the asymmetric 6-room


(a) $|\tilde{\mathscr{G}}| = 4$ **(b)** $|\tilde{\mathscr{G}}| = 5$

(c) $|\tilde{\mathscr{G}}| = 6$ **(d)** $|\tilde{\mathscr{G}}| = 7$

Figure 6.11: Goal clusters induced by the bottleneck $\tilde{G}_2$ on the secondary, state-information modulating goal-information pathway in a 9-room grid world navigation task. Figures (a) to (d) show the mappings for increasing cardinality of the bottleneck variable.

world, the peak does not lie at a cluster number of 6, but it turns out that a mapping with one extra cluster leads to more effectiveness. This clustering splits (not shown) the large north-eastern room into two parts, such that the sizes of the individual clusters are more uniform.

6.4.2 Observation 5: Global Reference Frame

Finally, we perform an analysis analogous to observation 4 of the previous section, on the secondary goal-information pathway, constrained by the bottleneck $\tilde{G}_2$. Again, we achieve a mapping $p(\tilde{g}_2|g)$ that effects a factorisation of goals. The result of this analysis is shown in Fig. 6.11. Similar to the clusterings of the previous section, the world is divided into several regions, however note that in this case the regions are not always constrained by walls: clusters tend to 'spill' over into neighbouring rooms. Instead of reconstructing the local connectivity of cells, these regions seem to adhere to a more general notion of ‘nearness’ of cells, which transcends the walls of the world as if they were not present. Furthermore, the clusters are roughly evenly distributed around the centre of the environment, with small clusters marking the centre of the environment when the cardinality is high enough. Due to these properties, we interpret this abstraction as capturing the *global* relative placing of the goal cells regardless of *local* structure created by walls, a pattern that is reproduced robustly in other scenarios shown in this paper.

We can understand this organisation by considering what aspect of a state is important in selecting actions in goal-directed behaviour. We stress that state knowledge is only informative when considered with respect to the current goal: only by knowing how it relates to the goal can an agent decide which action to take. For example, when taking the stairs, one needs to consider where the current floor is relative to the target floor to decide whether to take the flight of stairs up or down.

The optimal clustering in the context of our secondary bottleneck, is then one that preserves the possibility to determine such relations as much as possible. In other words, the clustering must preserve a global *frame of reference* in which state information can be considered. From Fig. 6.11 we can see that such a clustering results in a decomposition of the environment into its most salient global directions (or bearings), while disregarding disruptions of the structure by local obstacles.

It is important to note that the global relations between states and goals is strongly determined by the set of available actions. Consider for instance the subset of ’north-eastern goals’, i.e. those shaded darkest in Fig. 6.11a. Knowing that the goal is in this subset allows the agent to use state knowledge to make the informative distinction of whether the goal is likely to the north or to the west. But this distinction is only relevant because the agent has access to distinct actions that define these directions. Differently defined actions would induce other relations with goals, and likely a different abstraction would appear as the goal-based frame of reference. In the extreme, a much less structured set of actions, e.g. one where the action has different results in different situations [75], most likely makes it difficult to construct a useful abstraction under constraints on the informational pathways.

6.5 Discussion

In this chapter I have studied goal-directed behaviour in the light of informational constraints on the generation of behaviour. Using the Relevant-Information framework, I have studied lower bounds on how strongly a goal-information working memory can be constrained to still allow for successful goal-directed behaviour. Furthermore, I have applied the variants of the Information Bottleneck method to study explicit constraints on goal-information pathways.

The various results of this inquiry complement the main story line that is arising: informational limitations induce an organisation on the internal decision making processes and the thereby produced behaviour. Furthermore, this organisation aligns with and captures the organisation of the environment and the set of possible goals: we observed 1) ritualised behaviour along preferred paths between salient transition points, 2) that these extrinsic transition points are represented intrinsically by large informational transitions, and 3) natural abstractions that factorise the environment into sub-components most fittingly given a restricted information pathway.

As in the previous chapter, it must be stressed that all these effects arise solely as direct results from considering informationally bounded cognition. This is in contrast to approaches where these features are formulated a-priori and explicitly searched for, such as in finding sub-goals [68, 10, 45, 88], state and goal abstractions [46, 64] and goal-clustering [27] in transfer-based reinforcement learning [101]. The results presented here propose an intrinsically grounded justification for defining sub-goals and other abstracted concepts of typical tasks, which I will discuss in more detail in Chap. 6.5.

Finally, these methods might also offer opportunities to shed light on empirical results on the organisation of cognition in biological organisms; the large information 'unloading' transitions found at the doorways in Fig. 6.7b could for instance be related to recent results that show how people forget information when walking through a door [79].

6.6 Methodological Details

6.6.1 Goal Information Transition Sampling

The following sampling method was used to find $I(G; A_t|\mathbf{E}_{t-1}, w_t)$ and obtain Fig. 6.7a. The mutual information between the goal and action is equal to the reduction in entropy of A_t resulting from knowing the value of G:

$$I(G; A_t|\mathbf{E}_{t-1}, W_t) = H(A_t|\mathbf{E}_{t-1}, w_t) - H(A_t|G, \mathbf{E}_{t-1}, w_t) \tag{6.6}$$

Because of the symmetry of information, it also holds that:

$$I(G; A_t|\mathbf{E}_{t-1}, w_t) = H(G|\mathbf{E}_{t-1}, w_t) - H(G|A_t, \mathbf{E}_{t-1}, w_t) \tag{6.7}$$

Considering the asymptotic equipartition theorem [31], these entropy terms can be estimated by drawing n i.i.d. samples from the combination of G, $\mathbf{E}_{t-1}$ and A_t, according to:

$$H(G|\mathbf{E}_{t-1}, w_t) \approx -\frac{1}{n}\sum_{i=1}^{n}\sum_{g_i} \log p(g_i|\mathbf{e}_{t-1,i}, w_t) \tag{6.8}$$

$$H(G|A_t, \mathbf{E}_{t-1}, w_t) \approx -\frac{1}{n}\sum_{i=1}^{n}\sum_{g_i} \log p(g_i|a_t, \mathbf{e}_{t-1,i}, w_t) \tag{6.9}$$

for large values of n. The probability distribution $p(g_i|\mathbf{e}_{t-1,i}, w_t)$ is obtained by applying a series of Bayesian updates $p(g_i|\mathbf{e}_k) \leftarrow \frac{1}{\mathcal{Z}} p(a_k|w_k, g_i)p(g_i|\mathbf{e}_{k-1})$ for k from 0 to t, where $\mathcal{Z}$ is a normalisation factor and $p(g_i|\mathbf{e}_{-1})$ is set to be uniform. Noting that $p(g_i|a_t, \mathbf{e}_{t-1,i}, w_t) = p(g_i|\mathbf{e}_{t,i}, w_t)$ gives the other distribution needed to obtain the approximation.

The estimation of $I(G; \mathbf{E}_{t-1}|A_t, w_t)$ is obtained in a similarly fashion.

6.6.2 RI and IB Self-consistent Solutions

The following subsections derive the self-consistent solutions of the relevant information and bottle-neck type problems discussed in the text, by determining the partial derivative of the respective Lagrangian and finding the zero of that gradient.

Relevant State Information

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(a_t|w_t,g),\beta\Big) = I(W_t;A_t|G) - \beta E[U_G^\pi(W_t,A_t)], \tag{6.10}$$

and its partial derivative with respect to $p(a_t|w_t,g)$:

$$\frac{\partial}{\partial p(a_t|w_t,g)}\Lambda\Big(\pi(a_t|w_t,g),\beta\Big) = p(w_t,g)\log\frac{p(a_t|w_t,g)}{p(a_t|g)} - p(w_t,g)\beta U_g^\pi(w_t,a_t) \tag{6.11}$$

Equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(a_t|w_t,g) = \frac{1}{\mathcal{Z}}p(a_t|g)\exp\Big[-\beta U_g^\pi(w_t,a_t)\Big], \tag{6.12}$$

where $\mathcal{Z}$ is a normalisation term.

Relevant Goal Information

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(a_t|w_t,g),\beta\Big) = I(G;A_t|W_t) - \beta E[U_G^\pi(W_t,A_t)], \tag{6.13}$$

and its partial derivative with respect to $p(a_t|w_t,g)$:

$$\frac{\partial}{\partial p(a_t|w_t,g)}\Lambda\Big(\pi(a_t|w_t,g),\beta\Big) = p(w_t,g)\log\frac{p(a_t|w_t,g)}{p(a_t|w_t)} - p(w_t,g)\beta U_g^\pi(w_t,a_t) \tag{6.14}$$

Equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(a_t|w_t,g) = \frac{1}{\mathcal{Z}}p(a_t|w_t)\exp\Big[-\beta U_g^\pi(w_t,a_t)\Big], \tag{6.15}$$

where $\mathcal{Z}$ is a normalisation term.

State-Goal Information Trade-Off

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(a_t|w_t,g,\beta)\Big) = \alpha I(W_t;A_t|G) + (1-\alpha)I(G;A_t|W_t) - \beta E[U_G^\pi(W_t,A_t)], \tag{6.16}$$

and its partial derivative with respect to $p(a_t|w_t,g)$:

$$\begin{aligned}\frac{\partial}{\partial p(a_t|w_t,g)}\Lambda\Big(\pi(a_t|w_t,g),\beta\Big) = \alpha p(w_t,g)\log\frac{p(a_t|w_t,g)}{p(a_t|w_t)} &+ \\ (1-\alpha)p(w_t,g)\log\frac{p(a_t|w_t,g)}{p(a_t|w_t)} &- \\ p(w_t,g)\beta U_g^\pi(w_t,a_t) &\end{aligned} \tag{6.17}$$

Equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(a_t|w_t,g) = \frac{1}{\mathcal{Z}}p(a_t|w_t)^\alpha p(a_t|g)^{1-\alpha}\exp\Big[-\beta U_g^\pi(w_t,a_t)\Big], \tag{6.18}$$

where $\mathcal{Z}$ is a normalisation term.

6.6.3 Goal Information Bottleneck

For G'_1, the Langrangian is given as:

$$\Lambda\Big(\pi(g'_1|g),\beta\Big) = I(G;G'_1) - \beta I(G'_1;A_t|W_t). \tag{6.19}$$

Taking its partial derivative with respect to $p(g'_1|g)$, equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(g'_1|g) = \frac{1}{\mathcal{Z}} p(g'_1) \exp\Big[\sum_{w_t} p(w_t) D_{KL}\Big(p(a_t|g,w_t)||p(a_t|g'_1,w_t)\Big)\Big], \tag{6.20}$$

where $\mathcal{Z}$ is a normalisation term, and the Kullback-Leibler divergence $D_{KL}\Big(p(a_t|g,w_t)||p(a_t|g'_1,w_t)\Big) = \sum_{a_t} p(a_t|g,w_t) \log \frac{p(a_t|g,w_t)}{p(a_t|g'_1,w_t)}$.

For G'_2, the Langrangian is given as:

$$\begin{aligned}\Lambda\Big(\pi(g'_2|g),\beta\Big) &= I(G;G'_2) - \beta I(W_t;A_t|G'_2) && (6.21)\\ &= I(G;G'_2) - \beta\Big(I(\{W_t,G'_2\};A_t) - I(G'_2,A_t)\Big). && (6.22)\end{aligned}$$

Taking its partial derivative with respect to $p(g'_2|g)$, equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$\begin{aligned}p(g'_2|g) = \frac{1}{\mathcal{Z}} p(g'_2) \exp\Big[\sum_{w_t} p(w_t) D_{KL}\Big(p(a_t|g,w_t)||p(a_t|g'_2,w_t)\Big) - \\ D_{KL}\Big(p(a_t|g)||p(a_t|g'_2)\Big)\Big]. \end{aligned} \tag{6.23}$$

Chapter 7

Sensor Evolution

> “To suppose that the eye, with all its inimitable contrivances for adjusting the focus to different distances, for admitting different amounts of light, and for the correction of spherical and chromatic aberration, could have been formed by natural selection, seems, I freely confess, absurd in the highest possible degree.[1]”
>
> *Charles Darwin*

7.1 Introduction

Figure 7.1, reproduced from the introduction, shows the hypothesised pressures and the trade-off that are the foundation of the work in this thesis. Recall that, if an organism does not operate at an optimal trade-off level, this assumes that there is a drive to increase fitness through more effective actuation processes that utilise the superfluous cognitive capacity, while another pressure pushes towards degeneration of the sensory and cognitive capabilities to be more efficient and do away with unneeded energy consumption. This implies an ‘arms race’ of sorts between an agent’s cognitive and behavioural facilities, which continues until the drives meet in the middle.

At this point a so called ‘Pareto-efficient’ optimum is reached, where a unilateral change in a single component will push the organism away from the optimal trade-off. With none of the pressures pushing along the curve, how may it be possible for evolution to still move up or down the curve?

Moving down the curve can be imagined as following a 2-step pattern: first, if the performance pressure drops, an organism is free to move down in the trade-off plane. This could be caused by the decrease, or disappearance of some environmental factor, such as a competitor or a predator. Secondly, if this does happen, parsimony will push it back, to a point lower on the optimal curve to where it was before.

Moving up the curve however seems more problematic. The parsimony pressure does not suddenly drop, cognitive burden always comes with a cost. A step against the parsimony pressure therefore will always result in a drive back towards the same point on

[1] Continued: “Yet reason tells me, that [...] the difficulty of believing that a perfect and complex eye could be formed by natural selection, though insuperable by our imagination, can hardly be considered real”

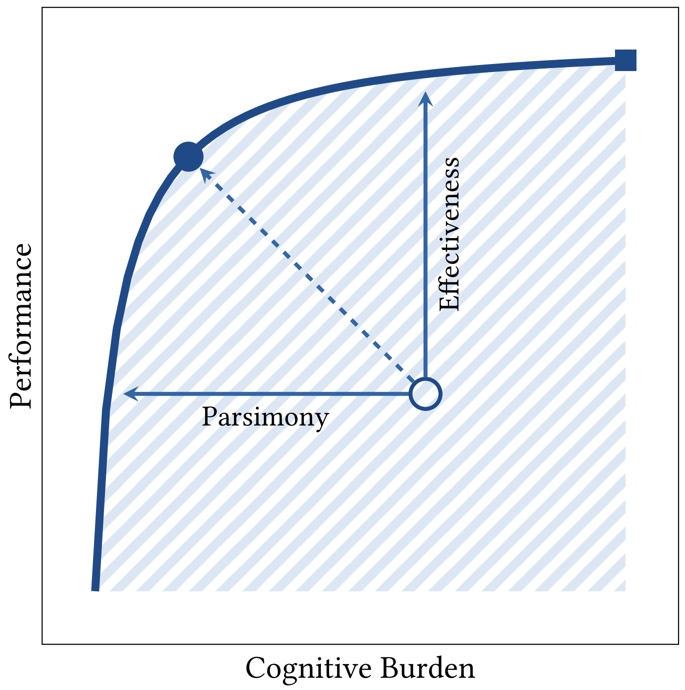


Figure 7.1: Trade-off between cognitive burden and behavioural performance, and parsimony and effectiveness pressures that form a drive towards the curve.

the curve. Moving from one point on the trade-off curve to one further up would thus need concurrent, well matched evolutionary steps in both cognitive and actuation space. Such synchronous, mutually reinforcing steps are highly unlikely, since in a random evolutionary scenario this requires two coordinated mutations. If this reasoning is correct, evolution would be slowed down considerably once a species' sensory-motor system has reached and operates on the optimal trade-off curve.

It is clear from nature however that this is not the case: species evolve continuously, and sometimes at considerable speeds. Species that are optimally adapted to a specific niche still seem able to rapidly specialise for and occupy another niche if the opportunity arises. Even more fascinating is that biological organisms do not seem to evolve simply towards any random locally optimal trade-off, but are instead driven to the *near-global* optima where their sensory capabilities are only limited by the laws of physics, some striking examples of which were given in the first chapter [15, 23, 33].

These considerations lead to the following questions. Firstly, how is it possible that species can evolve quickly from one local optimum to another, while local changes seemingly can only reduce their fitness, without the need of highly unlikely large and coordinated mutations? Secondly, what are possible factors that drive and facilitate evolution towards the ultimate limit of cognitive precision?

In the current chapter I will study these questions, specifically focusing on sensory bandwidth as an aspect of the cognitive burden. To help gain insight into these problems I will extend the information-theoretic framework build up earlier. Doing so, I will show 1) how the apparent co-dependence of sensory and actuation systems can be decoupled, 2) how this enables the gradual development of the combined system from one optimum

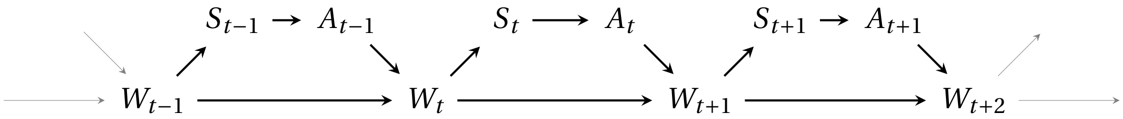


Figure 7.2: Causal Bayesian network of the perception-action loop used in this chapter. Note that the agent's, possibly incomplete, sensation of the world state is modelled explicitly this time.

to another, and 3) how this results in strong evolutionary pressure towards maximally advanced sensors.

In the following two sections I will introduce the formal frameworks that form the foundation of the approach. Next, I will develop a model of how the evolution of sensors and actuation can be uncoupled to facilitate transition from one locally optimal trade-off to another. Then, I will adapt this framework to model how evolution could drive sensors towards the upper limits of precision. Finally, I discuss fundamental information-theoretic properties of sensory systems that facilitate such processes, and argue that these properties constitute major, general, and fundamental drivers of sensor evolution.

7.2 Perception-Action Loop

As before, I will treat the perception-action loop as a causal Bayesian network, shown in Fig. 2.5b. In contrast to the previous chapters however, this time I will model the agent's sensor state at some time t explicitly, with the random variable S_t, which I will also refer to as the *sensor*. Values of this variable, specific instances of which are denoted with s_t are taken from the set $\mathscr{S}$. Note that this model results in a POMDP, where an agent may be able to observe only part of the world state.

The value of the agent's sensor is determined by the *sensor distribution*, which is the probabilistic mapping $p(s_t|w_t)$. The policy of the agent is then given by the conditional distribution $\pi(a_t|s_t)$, indicating that the agent acts upon the sensed state, not directly on the actual state.

7.3 Parsimonious Policy-Sensor Combination

In this framework, one can again ask for the minimal amount of informational burden required to achieve a fixed level of performance. There are now two channels that have informational burden: the sensory channel with average bandwidth $I(W_t; S_t)$, and the actuation channel with bandwidth $I(S_t; A_t)$. The parsimony hypothesis applies to both, so I will aim to find the optimal trade-off for each, starting with the latter, for reasons that will become clear after the following small proof.

Lemma 4.

$$I(S_t; A_t) = I(W_t; A_t)$$

Proof. The CBN of Fig. 7.2 imply the Markov chains $W_t \to S_t \to A_t$, and $A_t \to W_t \to S_t$, due to which $I(W_t;A_t|S_t) = I(S_t;A_t|W_t) = 0$. Thus:

$$\begin{aligned} I(S_t;A_t) &= I(S_t;A_t) + I(W_t;A_t|S_t) \\ &= I(W_t,S_t;A_t) \\ &= I(W_t;A_t) + I(S_t;A_t|W_t) \\ &= I(W_t;A_t) \end{aligned}$$

□

This means that one can first find the minimal actuation bandwidth independent of a sensor, by minimising $I(W_t;A_t)$ over all possible policies $\pi(a_t|w_t)$ (which I will denote a *direct* policy, as opposed to the definition of a policy above that selects an action based on the world state indirectly through a sensor). This minimum is found with the traditional relevant information method. In this chapter I will only treat full optimality, i.e with $\beta \to \infty$.

Once an *RI-optimal* direct policy is found, we can find a sensor mapping $p(s_t|w_t)$ for this policy. This is a mapping that minimises the bandwidth of the sensory channel, while still acquiring enough information to perform the minimal direct policy:

$$\min_{p(s_t|w_t)} I(W_t;S_t) \quad \text{subj. to} \quad I(S_t;A_t) = I(W_t;A_t) \tag{7.1}$$

This problem, in turn, is equivalent to a standard single-variate information bottleneck. The sensor mapping that is found is *minimally optimal*: it is optimal in the sense that it *retains all relevant information* to support a policy $\pi(a_t|s_t)$ that is consistent with the RI-optimal direct policy, and minimal in the sense that it *captures the minimum amount of information* about the world state to be able to reconstruct this information.

Once a sensor mapping is found, the final sensor based policy is given by:

$$\pi(a_t|s_t) = \sum_{w_t} \frac{1}{p(s_t)} \pi(a_t|w_t) p(s_t|w_t) p(w_t) \tag{7.2}$$

7.4 Uncoupled Sensor-Actuation Evolution

With the formal foundation in place, I will now develop an evolutionary model in which transitions between different locally optimal trade-offs are made feasible, by uncoupling the evolution of sensors and actuation.

In this model, we start out with an agent whose sensor and action selection mechanism operate on the optimal trade-off curve between informational burden and performance. Any point can be picked, but for now assume that the agent performs at the *globally* optimal trade-off at the end of the curve. This trade-off is fully determined by the utility of its actions and the world dynamics, and can be found using the RI and IB methods as discussed in the previous section. As noted before, this point seems to constitute an evolutionary dead-end, even more than any other locally, Pareto-optimal trade-off, since no improvement at all seems possible, which is problematic.

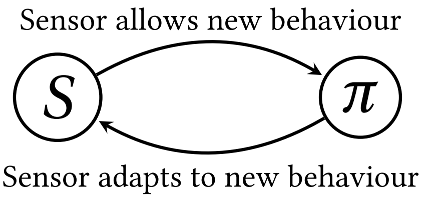


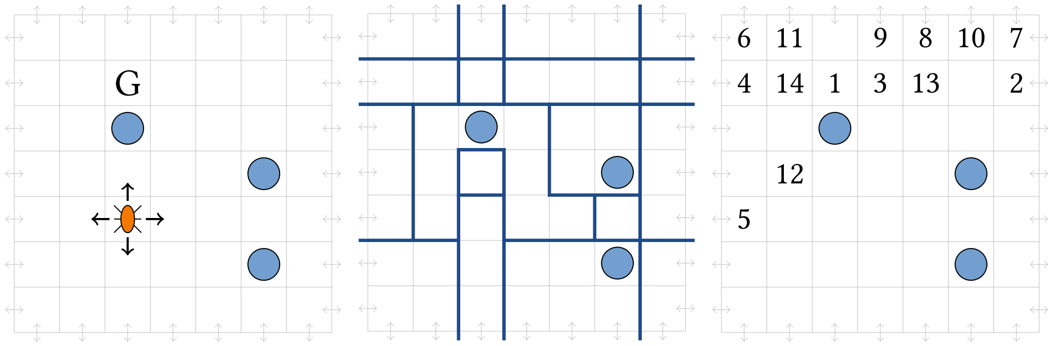
Figure 7.3: Graphical representation of uncoupled iterative evolution model

The solution to this problem is based on the idea that, given the currently evolved minimally optimal sensor, there could be *other niches* available for which this sensor is near-optimal. I will show that this view allows sufficient decoupling of the development of the components, which makes the necessary individual evolutionary steps much more likely.

The basic functioning of this model is visualised in Fig. 7.3: even when the sensor may be strictly minimal for a policy achieving optimal performance given one reward structure, this sensor may still give enough information to allow successful operation under a *different reward function*, and achievement of a *similar level of fitness* in this new scenario. In that case, evolution can drive the agent's behaviour, as expressed by its policy, to become optimal in this new situation, *without* the need of coordinated adaptation of the sensor. Once the transition to this new niche has started, the development of the sensor can instead *follow* that of the action selection mechanism, to again become minimally optimal. Here, I make no explicit assumption of what motivates such a transition between different niches, but possible drives may be toughening competition in the original niche, or perhaps simply evolutionary drift when the fitness achievable in both niches is similar enough.

To clarify this idea, I will apply this model to an example of the transformation of a sensor from nature. Tachinid flies posses a balloon-like sensor to detect movement of the head, which in the parasitoid *Therobia leonidei* has been evolved into an auditive sensor, which now is used in locating the bush-crickets that serve as its host [56]. This transformation can be explained in our model by noting that the original sensor, even if it would be fully optimised and minimal for its original use, may capture additional information that is relevant to the organism. In this case, the cognitive and actuation system of the organism can evolve to utilise this information, i.e. to better locate hosts, which constitutes the first step of the cycle above. Once this adaptation is set in motion, the evolution of the sensor can be driven towards higher auditive precision to better support the new strategy, which forms the second step of the cycle. These processes can then repeat until a new local optimum is reached, where the now auditive sensor is minimally optimal for its new function. Note that at no point of this process a coordinated adaptation of the combined sensory-actuation system is needed.

In this chapter, I will use a simple toroidal grid-world navigation task example, as depicted in Fig. 7.4, to show how this model works. The notion of different possible niches central to our model, formulated as different reward structures, is in such scenarios represented by a set of tasks, each with its according reward function. Here, each task is described by a goal state g that the agent needs to move into in as few steps as possible,



(a) (b) (c)

Figure 7.4: (a) Example 7 × 7 toroidal grid-world used to demonstrate the evolutionary model. The world-state consists of the agent's location. The agent receives a penalty of -1 for each step taken, unless it enters the goal state marked G, where reward is 0. The agent has access to 4 actions: move one cell north, east, south or west. Three randomly chosen cells, marked by grey disks, incur a reward of -5 when entered. (b) Location distinctions as given by minimally optimal sensor for task shown in (a). (c) Example of sequence of goals of first 12 tasks in expanding repertoire scenario.

Algorithm 3 Uncoupled Sensory-Motor Evolution

```
1: Select initial task $g$
2: Find RI-optimal direct policy $\pi_g(a_t|w_t)$
3: Use IB to find minimal optimal sensor $p(s_t|w_t)$ for this policy
4: Find the optimal policy $\pi_{g'}(a_t|s_t)$ for other tasks given current sensor
5: Determine task $g^*$ with highest performance given sensor, resolving ties by random
   selection
6: $g \leftarrow g*$
7: Repeat steps 2–3 for this new task
```

formalised by a reward function that penalises each step with a reward of -1, unless the agent enters the goal state, where the reward is 0. To prevent trivial solutions due to the high symmetry of the world, and to make lack of information about the world state more costly, several states are marked as 'danger' states that incur a cost of 5 upon entering.

A sensor in this world induces a possibly stochastic map, or clustering, of world states to a smaller set of sensor states, determining the precision in which the agent can observe its location. Because the sensor is achieved as an optimal trade-off with $\beta \to \infty$, the mapping ends up being deterministic [104]. Figure7.4b shows an example of a partitioning of the world by such a sensor.

The scenario, and the clustering induced by the sensor described here is similar to that of the multi-task scenario of the previous section. However the method for finding a policy is different, as well as how goals are selected, as described below. Furthermore, the clustering of the previous chapter was on *goal* states, whereas here the sensor clusters *world* states.

In this a scenario, one can formulate and perform the decoupled evolutionary itera-

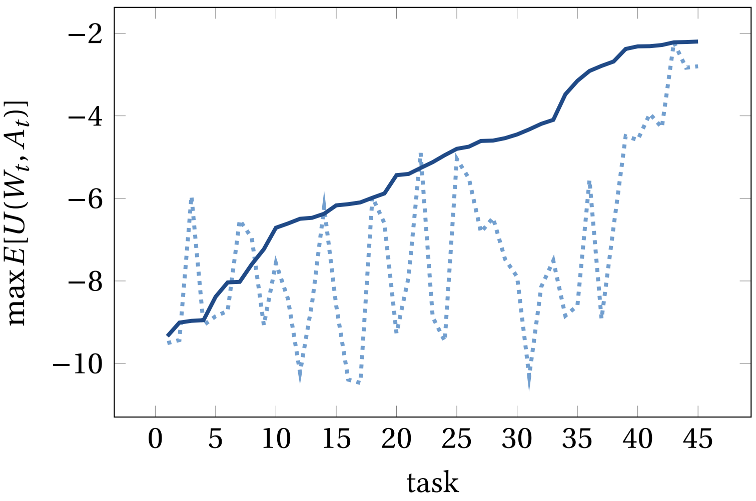


Figure 7.5: Typical example of utility achievable on each task using the minimal optimal sensor obtained for a specific initial task, denoted by the solid line, ordered from low to high achievable utility given this sensor. The task with the highest order number is the initial task for which the agent was optimised. The dashed line indicates the utility achievable using the action that would be taken for the initial task as the source of information, instead of the sensor input.

tions as follows. The agent starts with a single task selection g, which determines the current reward structure, or niche. Next we find the minimally optimal policy and sensor for this task. Finally, it is determined which other tasks can be performed well with the given sensor. The agent then moves onto the most rewarding of these tasks, and starts the optimisation process again. The technical description of this process is given in Alg. 3.

Step 4 of this algorithm, of which a detailed description can be found in the final section of this chapter, gives us a policy for all tasks, and the respective expected utility for these tasks under the policy that is found. The solid line in Fig. 7.5 shows a typical example of the result of plotting these maximally achievable utilities on the full range of tasks given the sensor for the initial task. The most striking observation in the context of the argument of this chapter, is that there is a group of tasks on which the agent can perform close to the optimum of the initial task, *despite* that the sensor that is used is fully optimised and minimised to provide *only* the information strictly relevant to the initial task.

When we obtain these results for all possible initial tasks, one can construct an evolutionary landscape, which indicates for which tasks an agent can still achieve near-optimal performance given the minimally optimal sensor of the predecessor task. This would show which evolutionary transitions are relatively easy to bring about, while at all times moving towards an optimal (local) information-utility trade-off, without the necessity of synchronised adaptation of both sensor and actuation, and with that how easy it is to move through the full space of niches. The relative ease of transitioning can for instance

	Diameter	Radius	Avg. Path Length	Connected
$\alpha = 0.9$	8	14	3.75	No
$\alpha = 1.0$	4	10	2.38	Yes
$\alpha = 1.1$	4	11	2.27	Yes
$\alpha = 1.5$	4	6	1.94	Yes

Table 7.1: Structural properties of the transition graphs of Fig. 7.6. Diameter: longest shortest path from one node to another; Radius: minimum of the longest distance from one node to another; Average path length: average over the shortest paths between all nodes; Connected: whether all nodes are (strongly) connected by some path with each other.

be determined by observing the ratio:

$$\frac{E_{g'}[U(S,A)|p_g(s_t|w_t)]}{E_g[U(S,A)|p_g(s_t|w_t)]}, \tag{7.3}$$

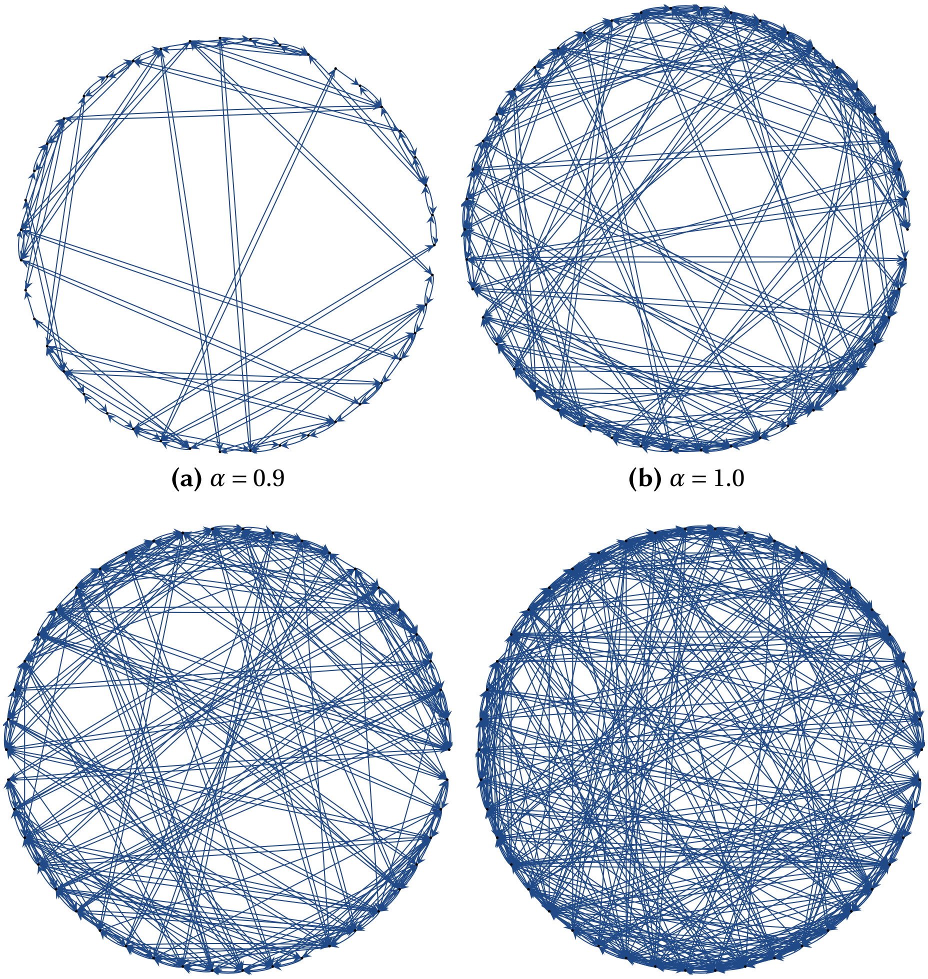
i.e. the ratio between the utility achievable on task g' given the sensor of initial task g, and the achievable utility on the initial task.

Figure 7.6 gives several examples of graphs obtained by connecting tasks for which the defined ratio is larger than some threshold $\frac{1}{\alpha}$. The inverse is used here, because the negative reward that is used is easier interpreted as a cost, such that the threshold α implies 'the expected cost on the next task can be no more than α times that of the initial task'. From these graphs different observations can be made. Firstly, Fig. 7.6a shows the striking possibility that there is a number of tasks for which the agent can achieve a *lower* expected cost than on a different task, with the minimally optimal sensor for this second task. Secondly, the connectivity of the graph grows rapidly when small increases in expected cost are tolerated, implying that in this scenario it is easy to move through the evolutionary landscape.

Such claims can be made more precise by applying the tools of graph theory to the results of this section. Table 7.1 for example lists some structural properties of the graphs of Fig. 7.6. These confirm the earlier claims: it is possible to move through all tasks without reducing expected utility at any step, and within on average 2.4 steps any point can be reached from any other, but in any case within 8 steps.

7.5 Sensor Evolution for Expanding Behaviour Repertoire

In the previous section I have given a model of how evolution could continuously drive an organism from being optimally adapted on one task (niche) to another. These steps can be seen as transitions from a point on the trade-off curve of one task to a point on the curve of another, and these transitions induce a drive to adapt a sensor for the new tasks. In this variant of the model, the complexity of the sensor could even decrease, if this precision is not necessary for the new task. Such an effect is seen in nature for instance


(a) $\alpha = 0.9$

(b) $\alpha = 1.0$

(c) $\alpha = 1.1$

(d) $\alpha = 1.5$

Figure 7.6: Directed graph showing feasible evolutionary transitions between different tasks under the uncoupled evolution model. Each task is represented by a point on the outer circle (in no particular order), and an arrow from one task to a second indicates that the minimally optimal sensor obtained for the first task allows an expected utility on the second task of no less than 95% than the maximum achievable on that task.

in blind Spalax mole rats and cave fish [37], that have occupied a niche where eyes are no longer relevant sensors and form an unnecessary burden. In this section I will show how the informational framework may increase understanding of how species could be driven towards the other, much more striking extreme that was noted in the introduction: where the sensory accuracy is pushed towards the limits of physics.

To do so, I change the interpretation of different reward functions from modelling specific mutually exclusive niches, only one of which an organism can occupy during its lifetime, to a set of goals that all can be imposed on an organism during its lifetime, drawn from some distribution $p(g)$. In this scenario, the overall performance of the agent is then determined by the expected utility averaged over *all* possible tasks, $E[U(S, A, G)]$. This means that there is a pressure to perform optimally on all tasks, instead of over-fitting on one or a small selection. This now gets closer to the model of the previous chapter, though now I aim to study the question that arose from that model: how can an agent acquire a repertoire of behaviour solving a set of tasks, while under the parsimony pressure?

The results of the previous section should supply a hint towards the answer to this question: even when minimally optimal on one task, it is still possible to perform other tasks at a high level of performance as well. The behavioural complexity could thus be increased by adding these other tasks to the agent's repertoire and retaining its existing skills, instead of doing away with them completely..

One can change the iterative decoupled evolutionary model of Alg. 3 at one point in order to fit this scenario: instead of letting the agent's sensor adapt fully to a new task, and by doing so move away from the old task, we let it adapt to incorporate the new task while *preserving* the optimality of its existing repertoire of behaviour. This means that, instead of adapting the agent's sensor to be optimal for the new task in step 3 of Alg. 3, we create an addition to the sensor, S'_t, that is optimised using an information bottleneck, such that it captures the relevant information for the new task, *beyond* what is already available in the existing sensor. Formally, this is done by solving:

$$\min_{p(s'_t|w_t)} I(W_t; S'_t) \quad \text{subj. to} \quad I(S_t, S'_t; A_t|G_k) \overset{!}{=} I(W_t; A_t|G_k), \tag{7.4}$$

where the value of G_k signifies a choice of task from the set of acquired tasks at iteration k of the algorithm.

This process can then be repeated, increasing the precision of the sensor at each step when necessary, until the agent's sensor has reached the maximum required precision to allow the agent to achieve all possible tasks optimally. This new iterative model is detailed in Alg. 4, of which step 5 is elaborated in the final section.

Performing this process in the grid-world scenario, and determining the overall performance of the agent at every iteration, gives the development curve shown in Fig. 7.7. This curve shows that indeed every adaptation to add a single task to the agent's repertoire monotonically increases the performance on the full range of tasks, even though at each step its sensor is *only* explicitly optimised to support a limited range of tasks. The most striking aspect however is how *rapidly* the sensor is driven toward the globally optimal precision: after optimisation for only 7 of the total of 46 tasks (less than 20%) the sensor is already precise enough to be able to perform near to optimal globally, with full

Algorithm 4 Sensor Evolution Towards Optimal Precision

```
1: Initialise ‘blind’ sensor ($|\mathscr{S}| = 1$)
2: Select initial task $g$
3: Find RI-optimal direct policy $\pi_g(a_t|w_t)$
4: Use IB to find minimal optimal addition to sensor $p(s'_t|w_t, s_t)$ for this policy
5: Combine the original sensor $S_t$ and the addition $S'_t$ into a new equivalent minimal
   sensor $S_t$
6: Find the optimal policy $\pi_{g'}(a_t|s_t)$ for other tasks given current sensor
7: Determine task $g^*$ with highest performance given sensor, resolving ties by random
   selection
8: $g \leftarrow g*$
9: Go to step 3 unless all tasks are treated
```

optimality possible after only 7 more epochs. Figure 7.4c shows the goals that were selected in these first 14 iterations. Note that the set of goals does not grow out from the first goal, but rather that successive goals can be some distance apart, but also that the final set of goals still only cover a distinct area, which apparently is enough to require a sensor to be accurate enough to reach any possible goal in the world optimally.

7.6 Concomitant Sensor Information as a Major Evolutionary Drive

The iterative model that I presented in this chapter is able to show that sensory evolution can be driven by the adoption of a novel behaviour/niche that is already well supported by the existing sensor, after which the sensor can be optimised for the new (repertoire of) behaviour. The results show that this process can rapidly bring about large evolutionary steps, based on the observation that, even when a sensor may be adapted fully for a single task, it still enables the achievement of different tasks near to optimality, or even fully optimally. An important question is whether this is an artefact of the particular examples chosen here, or of the constructed algorithms, or whether this is likely to hold more generally. In other words, are these dynamics generic? I argue that there is indeed a structural aspect of the PA-loop that facilitates adaptation towards novel optima, and that this aspect is reflected directly in the informational structure of the system.

From experience, in the information bottleneck paradigm it is common that the amount of information that a bottleneck variable (here: the sensor state) captures about the source variable (the world state) can be significantly larger than the amount it gives about the relevance variable (the action). Moreover, one can show formally that this inequality must hold for *all* possible combinations of worlds, sensors and policies, by employing the general information theoretic law of the data processing inequality [31]. In the framework of this chapter,this means that $I(W_t; S_t) \geq I(S_t; A_t)$, which one indeed encounters: in our scenarios the first term is between two to three times greater than the second. This observation is important: such a large amount of additional information available in the sensor state greatly increases the chance of a significant overlap with the information that is

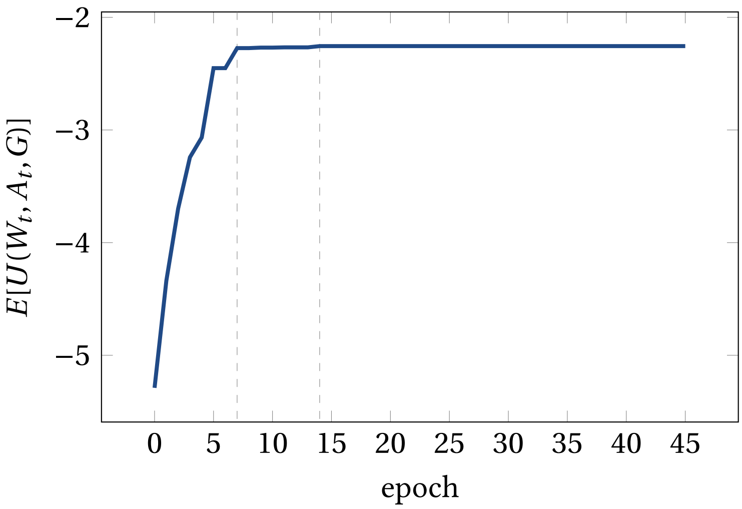


Figure 7.7: Typical example of the development curve of an agent in the grid-world navigation scenario, under the behaviour repertoire model. The curve plots the average over all tasks, after only being optimised for a subset of the tasks. The first dashed line shows where global optimality is approached closely, the second where full optimality has been achieved.

relevant for other task.

From this, I arrive at the following hypothesis:

Hypothesis 3. Concomitant sensor information*, the information that comes piggyback with the relevant information in a minimal optimal sensor, is a major factor in enabling sensory-actuation evolution.*

To test this, I will consider the maximum achievable performance on novel tasks using the sensor, which is likely to carry concomitant information, and compare it to the level achievable when strictly using only the minimum of information relevant to the initial task. This 'strict' relevant information is expressed in the final actions selected [83], so to obtain the latter performance one can alter step 4 of Alg. 3, to instead use the action selected according to the policy $\pi(a_t|w_t)$ as our 'sensor'. The results of this for the example grid-world scenario are depicted by the dashed curve in Fig. 7.5. They show that for many of the possible novel tasks, using the full sensor enables a significantly higher performance compared to utilising only the relevant information captured in the policy, as would be predicted from the above hypothesis.

7.7 Discussion

I have given a general model based on information-theoretical concepts of uncoupled sensor and actuation evolution, and shown how in this model evolutionary jumps between locally minimal optimal sensory-motor trade-offs can be facilitated.

The edges in a transition graph such as Fig. 7.6 give insight into the ease with which evolution can explore the full space of possibilities. Firstly, one can note that from each point a major subset of the other points can be reached through a limited number of transitions, implying that even a highly specialised species could evolve away into a wide range of completely different niches. Secondly, the fact that from many points not just one, but several points are directly reachable, indicates a possibility for diverging evolutionary pathways. And finally, the graph uncovers the irreversibility of parts of the evolutionary process. This is exhibited by a number of solutions that are only connected unidirectionally, indicating that the optimal sensor for one task is usable for the second, without the optimal sensor for the second supplying enough relevant information for the first task. Further graph-theoretical analysis of this graph, e.g. determining its radius, components, etc., or by integrating a similarity measure between tasks and/or between the minimally optimal sensors for those tasks, may uncover other interesting aspects, however this is outside the scope of the current chapter.

The most striking result of the current work is presented in Fig. 7.7, which shows a strong drive towards optimal sensory precision. The gradient of this curve indicates a significant pressure to optimise a sensor for novel behaviour. This occurs because this not only adapts the agent optimally to that specific novel behaviour, but the improvements of the sensor that follow this adaptation turn out to make a significant range of other beneficial behaviour feasible as well.

I argue again that the major facilitator of this process is the concomitant information, that is available in a sensor *beyond* that which is purely relevant, *even* in a sensor that is explicitly informationally minimal. Most notably, the presence of concomitant information is not an aspect of the specific model applied here, but derives from general, basic information-theoretical laws.

7.8 Methodological Details

7.8.1 Policy Optimisation for Novel Tasks

A *value-iteration* [97] type method is used to find the maximum achievable performance given a fixed sensor mapping $p(s_t|w_t)$. Here, the following is iterated until convergence, starting with a random policy π:

1. Perform value iteration until convergence of $U^\pi(w_t, a_t)$

2. Determine $U^\pi(s_t, a_t) = \sum_{w_t} p(w_t|s_t) U^\pi(w_t, a_t)$

3. Set policy to be greedy with respect to the new utility estimate, i.e. $\pi(a_t|s_t) \leftarrow 1/n$ if $U^\pi(s_t, a_t) = \max_{a'_t} U^\pi(s_t, a_t)$, otherwise $\pi(a_t|s_t) \leftarrow 0$. Here, n is the number of actions having the maximum utility, i.e. $|\{a_t : U^\pi(s_t, a_t) = \max_{a'_t} U^\pi(s_t, a_t)\}|$.

Finally, perform 1. to find the ultimate maximum performance $E[U^\pi(W_t, A_t)]$ given the final policy and sensor combination.

Due to the partial observability induced by a limited sensor, this process may not converge, but end up in an oscillation between a number of policies. In this case we stop after 1000 iterations and use the best policy in this oscillation. This may not be the global optimum, however this oscillation only occurs for tasks for which a sensor is notably unfitting, and thus does not influence our model, which is only concerned with well fitting tasks.

7.8.2 Sensor Extension and Merging

The bottleneck variables used in Algs. 3 and 4 (i.e. S_t and S'_t) have the same cardinality as the full world state variable, to ensure that there is no structural limitation on how much information they can capture. However, naively combining the existing sensor, S_t, and the addition optimised for a novel task, S'_t, in Alg. 4 leads to an exponential growth of the sensor size. As this makes the model computational unfeasible, and biologically implausible, we construct an equivalent minimal combination as follows (using Bayes' rule):

1. Determine $p(w_t|s_t, s'_t) = \frac{p(s_t|w_t)p(s'_t|w_t)p(w_t)}{p(s_t,s'_t)}$
2. Cluster all combinations s_t, s'_t that give sufficiently similar conditional distributions of W_t (as measured by the Jensen-Shannon divergence [31]) into a single new sensor state.

Practically, this results in a sensor with size no larger than that of the alphabet of world states.

Chapter 8

Discussion

This thesis started out with the model of behaviour generation and cognition repeated in Fig. 8.1. Throughout the following chapters the different parts of this model were studied by applying the same method on each:

1. Identify cognitive burden as the informational cost of some segment.
2. Determine how this cost trades off against some other cost, either in performance, or an additional informational cost.
3. Observe key structures, features and phenomena arising from this trade-off.

This one approach gave rise to a wide range of observations: preference for organized structure in action selection (Chap. 4); uncertainty aversion, active simplification of the environment, and development of a predictive memory (Chap. 5); ritualized behaviour, natural sub goals and abstractions (Chap. 6); the potentiality of rapid sensor evolution even under tight parsimony pressures (Chap. 7).

The most important aspect of these results is that all the phenomena appeared without an explicit requirement built into the scenarios: they came about due to self-organization under the parsimony pressure. One could definitely imagine other drives that give similar results. For instance, there would be an obvious pressure to develop a memory if it helps to get a better performance, by making available more relevant information that may not be directly available through direct sensing. Similarly, sub goals and abstractions may arise if they are needed to learn new tasks faster. From this point of view, if asking why we see such phenomena develop in nature, or why an artificial agent should develop them, the answer would be that they are needed to deal with some *external complexity*.

This is in stark contrast to the approach taken in this thesis, based on *internal simplicity*. It showed that the thinking described above may already be some steps too far: I explicitly ensured that none of the scenarios used required the agent to develop any of the observed phenomena; taking preferred paths, using memory, and finding natural transition points and abstractions did not make the agent perform his tasks any better.

This lays out a whole new possible story line for evolution and development: rather than having a species or an agent develop new structures or concepts when the need arises, we now have a fully intrinsic drive that could already give rise to them much earlier, which *then* could open up new capabilities and evolutionary or developmental pathways. A concurrent dependency similar to that discussed in Chap. 7 is broken by such a process: instead of requiring the development of new capabilities and the application of these on

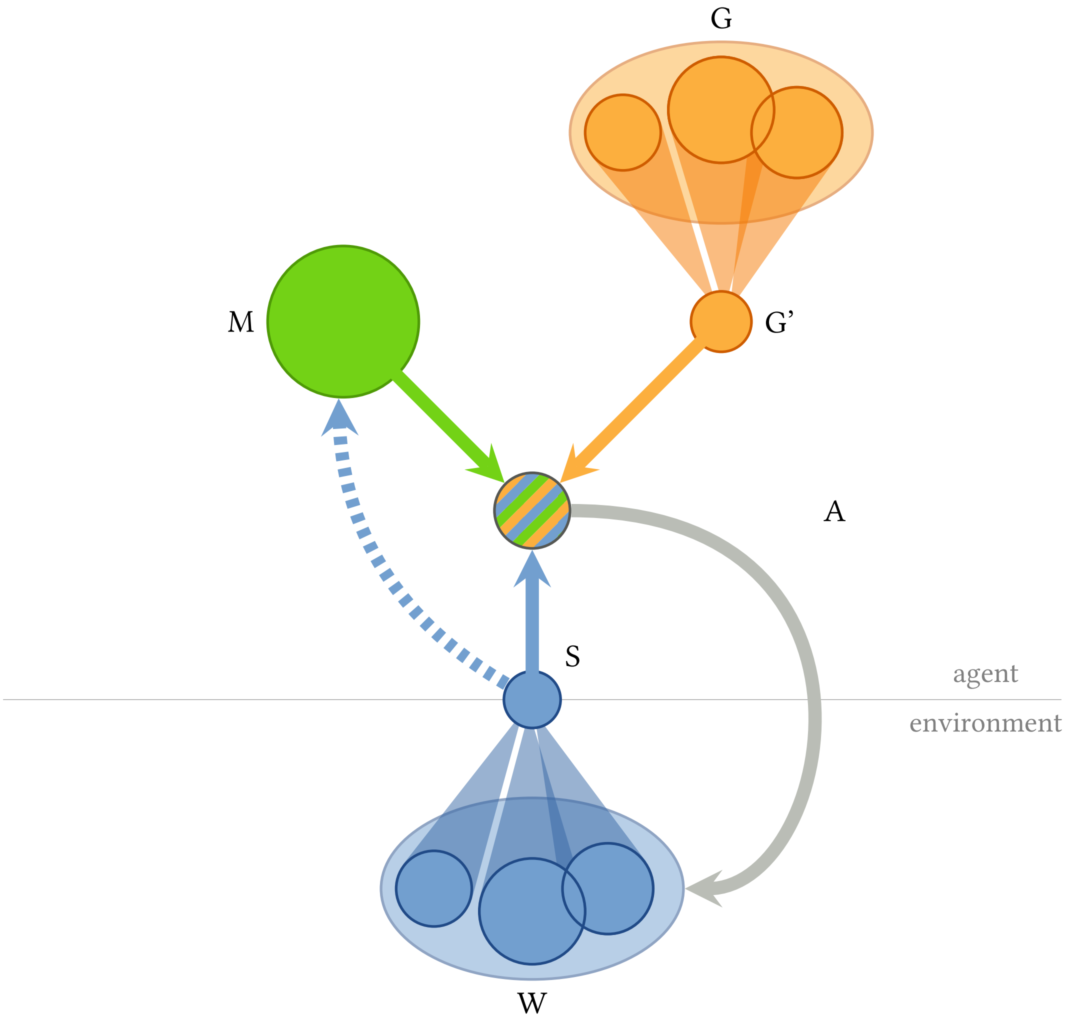


Figure 8.1: Sketch of the full cognitive model studied in this thesis.

novel things to happen together, the results of this thesis show the possibility of splitting these steps into an sequential, incremental process.

An important question is why the approach works. The model used was (purposefully) kept abstract, and the scenarios that were studied may be strong simplifications of real life. In the remainder of this final chapter I will discuss several possible extensions and open ends that may be studied to widen the approach and further test its results. My prediction however is that the encountered pattern, of self-organized structure under cognitive burden, will hold, and is fundamental, based on reasoning as follows.

All observed phenomena can be seen as reflecting salient features of the agent/environment system in the informational and behavioural outcome of the agent's decision making process. If cognitive processes are costly, which is difficult to deny they are, or more strongly, if they are under some finite limit, which is even more to be expected, an agent is *forced* to structure its decision making processes as effectively as possible to be able to act efficiently and successfully. I believe that such self-organisation of an agent can then *only* stem from the organisation that is available in the environment and the agent's goals and its embodiment. The agent can utilise this structure to alleviate its constraints, which results in this organisation being reflected internally.

These arguments and the obtained results, along with empirical evidence for informational constraints as a basis for the structuring of human behaviour [76], lead to the following proposal: a closed approach to generate a family of organisational concepts in a coherent way, by systematically applying cognitive constraints, as taken in this work in the form of information limitations, can constitute an important step towards guiding self-organisation. Put more simple: if you want to explain or achieve some organisation, simply assuming or imposing cognitive constraints may suffice.

Note that I have focused on informational constraints, but they may also have different forms, such as energy, time, or space constraints. Generally I would pose the following hypothesis as the main message of this thesis, and as an important possible foundation of future work:

Hypothesis 4. *Explicitly considering the properties of the cognitive facilities that a decision making process is embedded in, including but not limited to its informational aspects, should be a guiding principle: similar to how embodiment restricts and thereby shapes the interactions of an agent with the world, and* in combination with *embodiment, taking into account the agent's '*embrainment*' might be a key to understanding the organisation of cognition and behaviour.*

In the following section I will discuss in more detail how these ideas apply to, and how they may guide future work in a specific area of research: hierarchical decision making. The remainder of the chapter will discuss some further open avenues of research.

8.1 Hierarchy Through Parsimony

A popular model of decision making and behaviour generation, both in biology as in computational sciences, is that of a *decision hierarchy*. In such a model, sensory input is available at each layer in the decision making process; at each of these layers a decision is made in a different class, possibly based on different aspects of the sensory input. A hierarchy is formed through the *influence* that decisions at higher levels have on the decision making process at lower levels.

Models of decision hierarchies are often motivated by work from Tinbergen in the 1950s, on the behaviour of birds [103], and later refined by Baerends, a student of Tinbergen [9]. Some examples of where such hierarchies seem apparent are in vocal behaviour in singing birds [112] and ordering tasks in Capuchin monkeys [67]. Throughout decades much discussion was had about whether a decision hierarchy was a valid biological model, or a mere human projection on animal behaviour [32, 62, 63]. In computational applications this discussion was easier, and the application of hierarchical decision frameworks in artificial agents has flourished, initially in static systems [78], but more recently also in dynamically learning and organizing systems [13, 68, 25].

Now we know through empirical evidence that inherently hierarchical cognitive organizations are real-world entities (e.g in leeches [36], in mantids, cats crickets, and chickens [59], and in the human pre-frontal cortex [22]). In any case, at another scale the use of hierarchy is self evident: social animals for instance can form and operate under a hierarchical structure, such as a pack of wolves, and human government, armed forces, and businesses form hierarchies to control its more complex activities.

With all this evidence for hierarchies, one question is not yet fully answered: *why* would hierarchical organizations prevail? What is their benefit, and can we arrive at a quantitative theory of this benefit? Machine learning results have shown that *learning* with hierarchies can speed up the process. The results presented in this thesis could provide some insight into the benefits of *acting* with hierarchies.

Firstly, the layered action selection model with memory of Sec. 4.5 is one of the main frameworks used in the field of hierarchical reinforcement learning (HRL) [98, 13], from which the name 'option' for a higher level, temporally extended action is taken, with that difference that in the learning framework, the option choice is normally fixed over some time: a new option is only chosen at special states. These states can be seen as sub goals, which mark the end of a sub task that a specific option is said to perform.

From a learning point of view, the drive for building a behavioural hierarchy is to facilitate future learning: some options may be reused in new tasks, so the agent only has to learn when to start the option and then run its sub policy for several steps, instead of independently learning a separate policy for each separate state visited in those steps. However, the results of Chap. 4 indicate a more fundamental internal drive, active *before and regardless of* any need to learn different tasks: cognitive parsimony.

When at some point there is a drive towards hierarchy, which is now fully intrinsic, the next question that arises is which hierarchical structure is best, or most natural. The action selection hierarchy can be built to contain an almost arbitrary number of levels, where at each level options can be structured in many different ways. How many levels should there be, where should the boundary between them be, and what abstractions are best to make? From the point of view of learning, the answers to these questions should be guided by how well they facilitate future learning. However, this is problematic, because in general one can not expect which structure will perform well in future tasks: the whole problem is that the agent still has to learn about these tasks.

In this light it may not be surprising that commonly in the HRL literature, a hierarchy is built based on some heuristic which is assumed to provide useful abstractions. Often the abstraction is in the form of a sub goal state, that must be common for different tasks, and the heuristic is some feature of such states: times visited on successful performances [68, 96, 4], hand crafted abstractions [10], or various graph-theoretical ideas such as graph partitioning [89, 45] or betweenness [88].

The results of Chap. 6 indicate that again information minimization is enough to give rise to the sought after concepts, such as sub goals and abstractions, without the need to invoke learning capabilities. In their current form, these results don't fully answer the question of what is the best or most natural way to build a hierarchy (e.g. how many sub-goals are best, and can the natural abstractions be put in a natural hierarchy?), which is out of the scope of this thesis, but they are the first steps that open a path of future research that may lead there.

In the framework developed here, the most natural hierarchy is one for which some trade-off involving cognitive burden is the most favorable. We saw in the example of Chap. 4 that an additional layer reduced both the per-layer and the total sensory bandwidth, but one could point out different costs to trade-off against this reduction. For instance, with 4 options the maximum available bandwidth for the second level is 2 bits, but only just over half was used. Is there another number of layers where the bandwidth is more fitting, minimizing the Shannon redundancy, similar to how this allowed one to find a natural number of goal clusters in Sec. 6.4.1? Also, I have treated the lower levels as channels with side information, where this side information, the option choice at the next level, was freely available. However, we can imagine an additional channel that transmits this information, with its own additional bandwidth cost. One could then study the trade-off between this cost and the sensory bandwidth reduction, to find the informationally optimal hierarchical structure for the full range of relative costs, similar to what was done in Sec. 6.2.2 for the sensory versus goal information trade-off. Finally, one could combine the drives for cognitive parsimony and learning speed, to study whether these give rise to yet another trade-off.

8.2 Going Further

8.2.1 Completing the Cognitive Model

In this thesis, I have used the model drawn in Fig. 8.1. However, it was only possible to focus on the interaction of a few aspects of this model at a time. There are still several aspects that can be studied, and observing the model as a whole may uncover further interesting phenomena. As a matter of fact, one could point out several additional possible informational costs, or introduce further structure, and analyse many trade-offs that this brings about.

One example is some trade-off on the use of memory developed in Sec. 5.3. There, I only looked at how the memory can help reduce sensory burden, but it is likely that a memory adds its own cost. Actually, one could discern two types of costs: that of actually accessing the retained relevant information from the memory, and that of maintaining and/or possibly processing the information within the memory.

The first cost could readily be fit into the developed framework, by introducing $I(M; A)$ as another informational cost, with a new parameter to set different trade-offs between sensory and stored information, similar to how the sensory and goal relevant information were traded off in Sec. 6.2.2. With this extension, one can study under what trade-offs memory still develops, and how its structure may change under these trade-offs. One interesting outcome for instance could be the transition between memory at different timescales, which was obtained in a less principled manner by setting a hard limit on memory size in Sec. 5.3.3.

The second cost, that of storing and possibly processing of information within memory seems more difficult to define. One could add a cost in the form of $I(M; M')$, which may give the possibility to study the effects of memory size in a smoother way than simply setting its cardinality. However, there is one intuitive aspect of memory that such an informational cost does not capture: the dynamicity of memory, and its related cost. To explain, recall the memory experiment in Sec. 5.3.3, where the agent seemed to count either 1,2,1,2,... or 1,2,3,4,1,2,3,4,..., based on the taken pathway. It had to remember at which number it was, and dynamically update it at each step. However, there must also be some memory that is more static: throughout moving along a path, the agent should know on which path it is, to know up to what number to count. Intuitively, these two types of memory handle different types of information with different dynamics and possibly different costs. An interesting direction of future work could be to find a way to split these types and study their interactions: when do you need static memory, when dynamic; can one be traded off against the other? Answering these questions in the framework as constructed here seems difficult, and may require combining the relevant information idea with others, for instance time series complexity [84].

8.2.2 Continuity

The framework that was developed only treated scenarios having discrete time and state and action sets. This is not what one observes in real life; time and space appear continuous, and the possible ways of doing things nearly infinite. The main reason that the scope of this thesis was limited to discrete scenarios is practical: the information-theoretical tools and concepts that the study is based on are mature, practical, precise and well understood for these cases. In continuous scenarios, both the theory of Markov decision processes and information theory (with rate distortion theory specifically) only provide analytic solutions in rare, often trivial cases.

However, my believe is that the results in such discrete decision making processes are still valid, and that that they do indeed apply to actual cognitive processes. I would argue that in our experience, these processes are indeed of a more discrete nature: we tend to construct our models of the world from discrete entities, and think about discrete actions instead of continuous muscle control. To quote Barlow again [11]:

> […] the logic-like faculties of brains that lie behind higher mental functions are more interesting than the linear analysis that continuous signals link with most naturally, so my own bias is to regard the simpler concept of redundancy in discrete signals as more interesting and important.

More generally however, without invoking higher order faculties that may not be common in other species, the reasoning behind expecting the observation of internal reflection of, and self-organisation based on, salient features and structures in the environment as given above does not exclude continuous scenarios.

This is certainly not to say that it is not worthwhile to extend the framework to continuous scenarios: it would be an important step in developing it further, and test which of the observed phenomena are still seen in these cases, and what others may arise.

One consideration seemingly limits the applicability of the framework: although continuous variables have an infinite amount of possible values, only one may be the best choice. In a navigation scenario for example, the fastest path consists of a straight line to the target, compared to a grid world where there may be many paths that take the same amount of steps. Here one would already see one of the results of this thesis crumble: informational burden don't decide preferred paths anymore, only their optimality. As a matter of fact, it is even worse than that: the amount of information needed to select the one optimal value is infinite, so there seems no hope to reduce it for any of the channels that were considered.

However, luckily perfectly optimal solutions are hardly ever required. Though interesting from a learning theorist's point of view, a quick look at ourselves and nature indicates that it is often enough to be good enough, and no better, which is exactly in line with the original parsimony hypothesis. Reducing the pressure towards optimality makes more solutions good enough, which increases the range of possible choices and makes the amount of information that is required finite again.

8.2.3 Information Hiding and Multiple Agents

An important feature of information is that it is symmetric, formally given by the fact that $I(X;Y) = I(Y;X)$. In the context of acting agents, that means that an external observer, such as another agent, can obtain relevant information from observing actions of an agent, if there is overlap in what information is relevant for both. It can be shown that utilizing this preprocessed, or 'digested' information can motivate interesting social interactions, such as flocking [83]. The symmetry of information makes this possible without requiring that any agent knowingly transmits information to another agent.

From this point of view however, a relevant-information minimizing agent may be seen as anti-social: it seems to hide information from an external observer. A goal-information minimizing agent for instance only makes as little as possible known about its intentions.

This indicates a possibly very interesting opportunity for ongoing research: the further extension of the approach taken in this thesis into multi-agent scenarios. Some questions to study could be

- What trade-offs exist between egocentric information minimization and interacting with other agents, both adversarial, where information hiding may be beneficial, and social, where there may be pressures against parsimony?

- What other interactions may arise from egocentric parsimony? One imaginable possibility is the emergence of stigmergy from one agent reducing cognitive burden by offloading information into some change of the world, and another reducing his by using the digested information available in this change.

- What trade-offs would an explicit, active communication channel between agents and the associated burden of this channel bring about, and what structure would arise in this channel under the parsimony pressure?

8.2.4 Concomitant Information: a Universal Catalyst?

The previous chapter provided results that indicate that sensory evolution can be catalysed, even under strong evolutionary pressure to be informationally minimal, and introduced concomitant information as the main driver of this process. Again, it is instructive to note that the possible availability of this additional information is a direct result of the data processing inequality, one of the fundamental laws of information theory.

The fundamentality of this phenomenon leads me to hypothesise that it may not only be one of the major drives in sensor evolution, but that it could also play a large role in the evolution of many other aspects of cognitive systems. For instance, if the concomitant sensory information of the previous chapter is also relevant to future behaviour, it may significantly accelerate the evolution of memory.

Moreover, concomitant information is possible in any cognitive aspect where informational bottlenecks appear. Again taking memory as an example, one can consider the memory that arose from look-ahead information minimisation in Sec. 5.3: it formed an information bottleneck carrying predictive information about future states relevant to a single task. The concept of concomitant information would make the falsifiable prediction that this bottleneck may also carry additional information needed to ‘unlock’ the relevant information. This additional information could then act as a catalyst to adapt to different tasks, similar to the sensory case of the previous chapter.

Finally, taking this concept still further, it may even offer an insight into examples where relevant information happens to be captured by non-sensory systems, driving them to be adapted as useful sensors, as happened with lung-based hearing in amphibians [40]. Such directions of further exploration of the phenomena could give important insights into evolution and the importance of information therein.

Chapter A

Appendix

A.1 Publications

Peer-reviewed Work Contained in This Thesis

- Sander G. van Dijk, Daniel Polani, and Chrystopher L. Nehaniv. Hierarchical behaviours: Getting the most bang for your bit. In *Proc. European Conference on Artificial Life 2009, Budapest*, Budapest, Hungary, 2009.

- Sander G. van Dijk, Daniel Polani, and Chrystopher L. Nehaniv. What do you want to do today? relevant-information bookkeeping in goal-oriented behaviour. In Harold Fellermann, Mark Dörr, Martin Hanczyc, Lone L. Ladegaard, Sarah Maurer, Daniel Merkle, Pierre-Alain Monnard, Kasper Stø y, and Steen Rasmussen, editors, *Artificial Life XII: The 12th International Conference on the Synthesis and Simulation of Living Systems*, pages 176–183, Odense, Denmark, 2010. MIT Press.

- Sander G. van Dijk and Daniel Polani. Grounding subgoals in information transitions. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 105–111, Paris, France, 2011.

- Sander G. van Dijk and Daniel Polani. Look-ahead relevant information: Reducing cognitive burden over prolonged tasks. In *IEEE Symposium on Artificial Life*, pages 46–53, Paris, France, 2011.

- Sander G. van Dijk and D Polani. Informational drives for sensor evolution. In *Artificial Life XIII: The 13th International Conference on the Synthesis and Simulation of Living Systems*, pages 333–340, 2012.

- Sander G. van Dijk and Daniel Polani. Informational constraints-driven organization in goal-directed behavior. *Advances in Complex Systems*, 0(0):1350016, 2013. Electronic version of article published by ACS (c) copyright World Scientific Publishing Company.

Other Peer-Reviewed Publications

- Valerio Lattarulo and Sander G. van Dijk. Application of the “alliance algorithm” to energy constrained gait optimization. In *The 15th Annual RoboCup International Symposium*, pages 393–404, Istanbul, Turkey, 2011.

Other Publications

- Yo Sato, Ze Ji, and Sander G. van Dijk. I think i have heard that one before: Recurrence-based learning of word-like units – learning by 'echoing' in human-robot interaction–. In L. Gogate and G. Hollich, editors, *Theoretical and Computational Models of Word Learning: Implications for Psychology and Artificial Intelligence*, pages 327–349. IGI Global, 2013.

A.2 Equations, Derivations, and Solutions

A.2.1 Stationary State Distribution

For derivations and proofs of the results in this sub-section, the reader is referred to Norris [71].

Let $p(x_{t+1}|x_t)$ be a transition probability distribution that describes a first order, stationary Markov chain $(X_0, X_1, X_2, \ldots)$. A distribution $\bar{p}(x_t)$ is then said to be a stationary distribution for this chain if:

$$\bar{p}(x_{t+1}) = \sum_{x_t} \bar{p}(x_t) p(x_{t+1}|x_t), \tag{A.1}$$

i.e. if for some t the distribution of X_t is equal to $\bar{p}$, the distribution of all future state variables will be the same.

If the Markov chain is irreducible, the stationary distribution exists, and the system will converge onto this distribution after enough time. If one constructs the state-transition matrix $\mathbf{T}$, with its elements set as $p_{i,j} = p(x_{t+1} = j|x_t = i)$, and treat a distribution $p(x_t)$ as a vector $\mathbf{p}$, (A.1) can be stated as

$$\bar{\mathbf{p}} = \mathbf{T}\bar{\mathbf{p}}, \tag{A.2}$$

and the stationary distribution is found by applying $\mathbf{T}$ to any initial distribution until convergence:

$$\bar{\mathbf{p}} = \mathbf{T}^n \quad \text{as} \quad n \to \infty \tag{A.3}$$

A quicker way to find $\bar{\mathbf{p}}$ is by constructing a set of linear equations from (A.2), and the condition $\mathbf{1}^T\mathbf{p} = 1$:

$$\begin{bmatrix} (\mathbf{I} - \mathbf{T}) \\ \mathbf{1}^T \end{bmatrix} \bar{\mathbf{p}} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} \tag{A.4}$$

and solve these for $\bar{\mathbf{p}}$.

The specific transition probabilities that are used in this thesis are:

C-RI	$p(w_{t+1}\|w_t) = \sum_{a_t} \pi(a_t\|w_t) p(w_{t+1}\|w_t, a_t)$
LA-RI w mem.	$p(w_{t+1}, m_t\|w_t, m_{t-1}) = \sum_{j_t} \pi(j_t\|w_t, m_{t-1}) \mathbf{1}_{j_t}(m_t) p(w_{t+1}\|w_t, f_a(j_t))$
RGI	$p(w_{t+1}\|w_t) = \sum_{g,a_t} \pi(a_t\|w_t, g) p(w_{t+1}\|w_t, a_t) p(g)$

A.2.2 Information Derivatives

What follows are solutions to the derivatives of common information measures, used to derive the self-consistent solutions to the optimization problems of this thesis. All are based on the known equality:

$$\frac{\partial}{\partial x} f(x) \log g(x) = \left[\frac{\partial}{\partial x} f(x)\right] \log g(x) + \frac{f(x)}{g(x)} \left[\frac{\partial}{\partial x} g(x)\right]$$

Mutual Information 1

$$\begin{aligned}
\frac{\partial}{\partial p(\tilde{x}|x)} I(X;\tilde{X}) &= \frac{\partial}{\partial p(\tilde{x}|x)} \sum_{x,\tilde{x}} p(x,\tilde{x}) \log p(\tilde{x}|x) - \frac{\partial}{\partial p(\tilde{x}|x)} \sum_{x,\tilde{x}} p(x,\tilde{x}) \log p(\tilde{x}) \\
&= [p(x) \log p(\tilde{x}|x) + p(x)] - [p(x) \log p(\tilde{x}) + \sum_{x'} p(x'|\tilde{x}) p(x)] \\
&= p(x) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})}
\end{aligned}$$

Mutual Information 2

If the Markov property $\tilde{X} \to X \to Y$ holds, then

$$\begin{aligned}
\frac{\partial}{\partial p(\tilde{x}|x)} p(\tilde{x},y) &= \frac{\partial}{\partial p(\tilde{x}|x)} \sum_x p(\tilde{x}|x) p(y|x) p(x) = p(x,y) \\
\frac{\partial}{\partial p(\tilde{x}|x)} p(\tilde{x}|y) &= \frac{\partial}{\partial p(\tilde{x}|x)} \sum_x p(\tilde{x}|x) p(x|y) = p(x|y) \\
\frac{\partial}{\partial p(\tilde{x}|x)} p(\tilde{x}) &= p(x),
\end{aligned}$$

which gives:

$$\begin{aligned}
\frac{\partial}{\partial p(\tilde{x}|x)} I(\tilde{X};Y) &= \frac{\partial}{\partial p(\tilde{x}|x)} \sum_{\tilde{x},y} p(\tilde{x},y) \log p(\tilde{x}|y) - \frac{\partial}{\partial p(\tilde{x}|x)} \sum_{\tilde{x},y} p(\tilde{x},y) \log p(\tilde{x}) \\
&= \sum_y [p(x,y) \log p(\tilde{x}|y) + p(y) p(x|y)] - \sum_y [p(y,x) \log p(y) + p(y|\tilde{x}) p(x)] \\
&= p(x) \sum_y p(y|x) \log \frac{p(y|\tilde{x})}{p(y)}
\end{aligned}$$

Conditional Mutual Information 1

$$\begin{aligned}\frac{\partial}{\partial p(\tilde{x}|x,z)}p(\tilde{x}|z) &= \frac{\partial}{\partial p(\tilde{x}|x,z)}\sum_x p(\tilde{x}|x,z)p(x|z) = p(x|z)\\ \frac{\partial}{\partial p(\tilde{x}|x,z)}I(X;\tilde{X}|Z) &= \frac{\partial}{\partial p(\tilde{x}|x,z)}\sum_{x,\tilde{x},z} p(x,\tilde{x},z)\log p(\tilde{x}|x,z)-\\ &\quad\frac{\partial}{\partial p(\tilde{x}|x,z)}\sum_{x,\tilde{x},z} p(x,\tilde{x},z)\log p(\tilde{x}|z)\\ &= p(x,z)\log p(\tilde{x}|x,z)+p(x,z)-\\ &\quad p(x,z)\log p(\tilde{x}|z)-\sum_{x'}\frac{p(x',\tilde{x},z)}{p(\tilde{x}|z)}p(x|z)\\ &= p(x,z)\log\frac{p(\tilde{x}|x,z)}{p(\tilde{x}|z)}\end{aligned}$$

A.2.3 Conditional Mutual Information 2

If the Markov properties $\tilde{X}\to X\to Y$ and $\tilde{X}\to X\to Z$ hold, then

$$\begin{aligned}\frac{\partial}{\partial p(\tilde{x}|x)}p(\tilde{x},y,z) &= \frac{\partial}{\partial p(\tilde{x}|x)}\sum_x p(\tilde{x}|x)p(y|x,z)p(z|x)p(x) = p(x,y,z)\\ \frac{\partial}{\partial p(\tilde{x}|x)}p(\tilde{x}|y,z) &= \frac{\partial}{\partial p(\tilde{x}|x)}\sum_x p(\tilde{x}|x)p(x|y,z) = p(x|y,z)\\ \frac{\partial}{\partial p(\tilde{x}|x)}p(\tilde{x}|z) &= \frac{\partial}{\partial p(\tilde{x}|x)}\sum_x p(\tilde{x}|x)p(x|z) = p(x|z),\end{aligned}$$

which gives

$$\begin{aligned}\frac{\partial}{\partial p(\tilde{x}|x)}I(\tilde{X};Y|Z) &= \frac{\partial}{\partial p(\tilde{x}|x)}\sum_{\tilde{x},y,z}p(\tilde{x},y,z)\log p(\tilde{x}|y,z)-\frac{\partial}{\partial p(\tilde{x}|x)}\sum_{\tilde{x},y,z}p(\tilde{x},y,z)\log p(\tilde{x}|z)\\ &= \sum_{y,z}[p(x,y,z)\log p(\tilde{x}|y,z)+p(x,y,z)]-\\ &\quad\sum_{y,z}[p(x,y,z)\log p(\tilde{x}|z)+\frac{p(\tilde{x},y,z)}{p(\tilde{x}|z)}p(x|z)]\\ &= p(x)\sum_z p(z|x)\sum_y p(y|x,z)\log\frac{p(y|\tilde{x},z)}{p(y|z)}\end{aligned}$$

A.2.4 Conditional Mutual Information 3

$$\begin{aligned}\frac{\partial}{\partial p(\tilde{x}|x)}I(\tilde{X},Z;Y) &= \frac{\partial}{\partial p(\tilde{x}|x)}[I(\tilde{X};Y|Z)+I(Z;Y)]\\ &= \frac{\partial}{\partial p(\tilde{x}|x)}I(\tilde{X};Y|Z),\end{aligned}$$

which results in the solution of the previous section.

Bibliography

[1] Kevin R Abbott and Thomas N Sherratt. Optimal sampling and signal detection: unifying models of attention and speed–accuracy trade-offs. *Behavioral Ecology*, 24(3):605–616, 2013.

[2] Tom Anthony, Daniel Polani, and Chrystopher L Nehaniv. Impoverished empowerment : ‘ meaningful ’ action sequence generation through bandwidth limitation. In *Proc. European Conference on Artificial Life 2009, Budapest*, 2009.

[3] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *Information Theory, IEEE Transactions on*, 18(1):14–20, 1972.

[4] Mehran Asadi and Manfred Huber. Accelerating action dependent hierarchical reinforcement learning through autonomous subgoal discovery. In *In Proceedings of the ICML 2005 Workshop on Rich Representations for Reinforcement Learning*, 2005.

[5] Joseph J Attick. Could information theory provide an ecological theory of sensory processing? *Network*, pages 213–251, 1992.

[6] Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, May 1954.

[7] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich. Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B - Condensed Matter and Complex Systems*, 63(3):329–339, 2008.

[8] Nihat Ay, Holger Bernigau, Ralf Der, and Mikhail Prokopenko. Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theory in Biosciences*, 131:161–179, 2012.

[9] Gerald P. Baerends. The functional organization of behaviour. *Anuimal Behaviour*, 24:726–738, 1976.

[10] Bram Bakker and J. Schmidhuber. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In *Intelligent Autonomous Systems 8*, pages 438–445. IOS Press, 2004.

[11] H. Barlow. Redundancy reduction revisited. *Network*, 12:241–254, 2001.

[12] HB Barlow. Possible principles underlying the transformation of sensory messages sensory communication ed wa rosenblith. In W. Rosenblith, editor, *Sensory Communication*, volume 1, pages 217–234. MIT Press, Cambridge, MA, USA, 1961.

[13] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(1-2):41–77, 2003.

[14] Jonathan Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39, 1997.

[15] D. A. Baylor, T. D. Lamb, and K. W. Yau. Responses of retinal rods to single photons. *The Journal of physiology*, 288:613–634, March 1979.

[16] Baltasar Beferull-Lozano, Robert L. Konsbruck, and Martin Vetterli. Rate-distortion problem for physics based distributed sensing. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, IPSN '04, pages 330–339, New York, NY, USA, 2004. ACM.

[17] Richard Bellman. A markovian decision process. *Indiana University Mathematics Journal*, 6:679–684, 1957.

[18] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 3 edition, 2005.

[19] W Bialek, I Nemenman, and N Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–63, November 2001.

[20] Richard E Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18:460–473, 1972.

[21] Alexander Borst and Frédéric E. Theunissen. Information theory and coding. *Nature Neuroscience*, 2(11):947–957, 1999.

[22] Matthew M Botvinick. Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci. (Regul. Ed.)*, 12(5):201–208, 2008.

[23] T. H. Bullock and F. P. J. Diecke. Properties of an infra-red receptor. *J. Physiol.*, 134:47–87, 1956.

[24] Daniel a Butts. How much information is associated with a particular stimulus? *Network (Bristol, England)*, 14(2):177–87, May 2003.

[25] Özgür Şimşek and Andrew G. Barto. *Using relative novelty to identify useful temporal abstractions in reinforcement learning.* ACM Press, New York, New York, USA, July 2004.

[26] Philippe Capdepuy, Daniel Polani, and Chrystopher L. Nehaniv. Constructing the basic umwelt of artificial agents: An information-theoretic approach. In Fernando Almeida e. Costa, Luis Mateus Rocha, Ernersto Costa, Inman Harvey, and António Coutinho, editors, *Proceedings of the Ninth European Conference on Artificial Life*, pages 375–383. Springer, 2007.

[27] James L. Carroll and Kevin Seppi. Task similarity measures for transfer in reinforcement learning task libraries. In *The 2005 International Joint Conference on Neural Networks, (IJCNN 2005)*, pages 803–808, 2005.

[28] Philip A. Chou and Zhourong Miao. Rate-distortion optimized streaming of packetized media. *IEEE Transactions on Multimedia*, pages 390–404, 2006.

[29] Andy Clark and David J. Chalmers. The extended mind, 1998.

[30] Colin W. Clark and Reuven Dukas. The behavioral ecology of a cognitive constraint: limited attention. *Behavioral Ecology*, 14(2):151–156, 2003.

[31] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, January 1991.

[32] Richard Dawkins. Hierarchical organisation: A candidate principle for ethology. In P. P. G. Bateson and R. A. Hinde, editors, *Growing points in ethology*, pages 7–54. Cambridge University Press, 1976.

[33] Winfried Denk and Watt W. Webb. Thermal-noise-limited transduction observed in mechanosensory receptors of the inner ear. *Phys. Rev. Lett.*, 63:207–210, Jul 1989.

[34] Ralf Der, Frank Güttler, and Nihat Ay. Predictive information and emergent cooperativity in a chain of mobile robots. In *Artificial Life XI.* MIT Press, 2008.

[35] M R DeWeese and M Meister. How to measure the information gained from one symbol. *Network (Bristol, England)*, 10(4):325–40, November 1999.

[36] Teresa Esch and Wiliam B. Kristan Jr. Decision-making in the leech nervous system. *Integrative and Comparative Biology*, 42(4):716–724, 2002.

[37] Daniel Fong, Thomas Kane, and David Culver. Vestigialization and loss of nonfunctional characters. *Annual Review of Ecology & Systematics*, 26:249–268, 1995.

[38] Antonia F De C Hamilton and Scott T Grafton. Goal representation in human anterior intraparietal sulcus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(4):1133–7, January 2006.

[39] Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Physical Review E*, 87:012130, 2013.

[40] T. E. Hetherington and E. D. Lindquist. Lung-based hearing in an earless anuron amphibian. *Journal of Comparative Physiology*, 184:395–401, 1999.

[41] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[42] W R Jeffery. Cavefish as a model system in evolutionary developmental biology. *Developmental biology*, 231(1):1–12, March 2001.

[43] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[44] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principles of neural science.* McGraw-Hill, Health Professions Division, 2000.

[45] S. Kazemitabar and Hamid Beigy. Automatic discovery of subgoals in reinforcement learning using strongly connected components. *Advances in Neuro-Information Processing*, pages 829–834, 2009.

[46] Ghorban Kheradmandian and Mohammad Rahmati. Automatic abstraction in reinforcement learning using data mining techniques. *Robotics and Autonomous Systems*, 57(11):1119–1128, 2009.

[47] Young-Han Kim. A coding theorem for a class of stationary channels with feedback. *IEEE Transactions on Information Theory*, 54(4):1488–1499, 2008.

[48] K. Kirschfeld. The resolution of lens and compound eyes. In F. Zettler and R. Weiler, editors, *Neural principles in vision*, pages 354–370. Springer, 1976.

[49] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Representations of space and time in the maximization of information flow in the perception-action loop. *Neural computation*, 19(9):2387–432, September 2007.

[50] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Keep your options open: an information-based driving principle for sensorimotor systems. *PloS one*, 3(12):e4018, January 2008.

[51] A.S. Klyubin, D Polani, and C.L. Nehaniv. Organization of the information flow in the perception-action loop of evolved agents. In R. S. Zebulum, D. Gwaltney, G. Hornby, D. Keymeulen, and A. Lohn, J. andStoica, editors, *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, pages 177–180. Published by the IEEE Computer Society, 2004.

[52] A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: a universal agent-centric measure of control. *Procs of the 2005 IEEE Congress on Evolutionary Computation 1 pp.128 -*, 135:128–135, 2005.

[53] R. Kortmann, E. Postma, and J. van den Herik. Evolution of visual resolution constrained by a trade-off. *Artif Life*, 7(2):125–145, 2001.

[54] J.R. Krebs and N.B. Davies. *Introduction to Behavioural Ecology.* Blackwell Science Ltd, 1993.

[55] I Kupfermann, E.R. Kandel, and S. Iversen. Motivational and addictive states. In E.R. Kandel, J.H. Schwarz, and T.M. Jessell, editors, *Principles of neural science*, pages 998–1013. Elsevier, 2000.

[56] Reinhard Lakes-Harlan and Klaus-Gerhard Heller. Ultrasound-sensitive ears in a parasitoid fly. *Naturwissenschaften*, 79:224–226, 1992.

[57] S B Laughlin, R R de Ruyter van Steveninck, and J C Anderson. The metabolic cost of neural information. *Nature neuroscience*, 1(1):36–41, May 1998.

[58] Jonathan Lipton, Gunnar Kleemann, Rajarshi Ghosh, Robyn Lints, and Scott W. Emmons. Mate searching in caenorhabditis elegans: A genetic model for sex drive in a simple invertebrate. *The Journal of Neuroscience*, 24(34):7427–7434, 2004.

[59] Eckehard Liske. The hierarchical organization of mantid behavior. In Frederick R. Prete, Harrington Wells, Patrick H. Wells, and Lawrence E. Hurd, editors, *The Praying Mantids*, pages 224–250. Johns Hopkins University Press, 1999.

[60] Michael L Littman. A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125, 2009.

[61] Max Lungarella and O. Sporns. Information self-structuring: Key principle for learning and development. In *Proc. of the 4 th Int. Conf. on Development and Learning*, pages 25–30. IEEE, 2005.

[62] Pattie Maes. A bottom-up mechanism for behavior-selection in an artificial creature. In J.-A. Meyer and S.-W Wilson, editors, *From Animals to Animats*, pages 169–175. MIT Press/Bradford Books, 1991.

[63] Pattie Maes. Modeling adaptive autonomous agents. *Artificial Life*, 1(1-2):135–162, January 1993.

[64] Shie Mannor, Ishai Menache, Amit Hoze, and Uri Klein. Dynamic abstraction in reinforcement learning via clustering. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 71, New York, NY, USA, 2004. ACM.

[65] Andrew C. Mason, Michael Oshinsky, and Ron Hoy. Hyperacute directional hearing in a microscale auditory system. *Nature*, 410:686–690, April 2001.

[66] J. Massey. Causality, feedback and directed information. In *Proc. 1990 Intl. Symp. on Info. Th. and its Applications*, pages 27–30. Citeseer, 1990.

[67] B McGonigle. Concurrent disjoint and reciprocal classification by cebus apella in seriation tasks: Evidence for hierarchical organization. *Animal cognition*, 2003.

[68] A. McGovern and A.G. Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *ICML ´01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 361–368, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[69] P Read Montague, Steven E Hyman, and Jonathan D Cohen. Computational roles for dopamine in behavioural control. *Nature*, 431(7010):760–7, October 2004.

[70] E. A. Newman and P. H. Hartline. The infrared "vision" of snakes. *Scientific American*, 246(3):98–107, 1982.

[71] James R. Norris. *Markov Chains.* Cambridge Series in Statistical and Probabilistic Mathematics (no 2). Cambridge University Press, 1998.

[72] Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1):1–4, 2003.

[73] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[74] Rolf Pfeifer, Max Lungarella, Olaf Sporns, and Yasuo Kuniyoshi. On the information-theoretic implications of embodiment – principles and methods. In *50 Years of Artificial Intelligence*, volume 4850, pages 76–86. Springer-Verlag, 2007.

[75] D. Polani. An informational perspective on how the embodiment can relieve cognitive burden. In *IEEE Symposium on Artificial Life*, pages 78 – 85, 2011.

[76] Daniel Polani. Information: currency of life? *HFSP journal*, 3(5):307–16, October 2009.

[77] Daniel Polani, C. Nehaniv, Thomas Martinetz, and J.T. Kim. Relevant information in optimized persistence vs. progeny strategies. In *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, pages 337–343. Citeseer, 2006.

[78] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[79] Gabriel A Radvansky, Sabine A Krawietz, and Andrea K Tamplin. Walking through doorways causes forgetting: Further explorations. *Quarterly journal of experimental psychology*, 64(8):1632–45, August 2011.

[80] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition).* Prentice Hall, December 2002.

[81] Marco Saerens, Youssef Achbany, François Fouss, and Luh Yen. Randomized shortest-path problems: Two related models. *Neural Comput.*, 21(8):2363–2404, 2009.

[82] Naohiro Saito, Hajime Mushiake, Kazuhiro Sakamoto, Yasuto Itoyama, and Jun Tanji. Representation of immediate and final behavioral goals in the monkey prefrontal cortex during an instructed delay period. *Cerebral cortex (New York, N.Y. : 1991)*, 15(10):1535–46, October 2005.

[83] Christoph Salge and Daniel Polani. Digested information as an information theoretic motivation for social interaction. *Journal of Artificial Societies and Social Simulation*, 14(5), January 2010.

[84] Cosma Rohilla Shalizi. *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata.* PhD thesis, 2001.

[85] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. In *Algorithmic Learning Theory*, pages 92–107. Springer, 2008.

[86] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[87] Claude. E. Shannon. The zero-error capacity of a noisy channel. *IRE Transactions on Information Theory*, 1956.

[88] Özgür Şimşek and Andrew G. Barto. Skill characterization based on betweenness. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1497–1504, 2009.

[89] Özgür Şimşek, A.P. Wolfe, and A.G. Barto. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd international conference on Machine learning*, pages 816–823. ACM, 2005.

[90] Sinan Sinanovic and Don H. Johnson. Toward a theory of information processing. *Signal Processing*, 87:1326–1344, 2007.

[91] Noam Slonim. *The Information Bottleneck: Theory and Applications.* PhD thesis, Hebrew University, 2002.

[92] Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate information bottleneck. *Neural computation*, 18(8):1739–89, August 2006.

[93] Olaf Sporns and Max Lungarella. Evolving coordinated behavior by maximizing information structure. In *Artificial life X: proceedings of the tenth international conference on the simulation and synthesis of living systems*, pages 323–329. Citeseer, 2006.

[94] Susanne Still. Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85:28005, 2009.

[95] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 2011. To appear.

[96] Martin Stolle and Doina Precup. Learning options in reinforcement learning. *Lecture Notes in Computer Science*, 2371:212–223, 2002.

[97] Richard S Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, USA, March 1998.

[98] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

[99] S. Tatikonda. The sequential rate distortion function and joint source-channel codeing with feedback. In *41st Allerton Conference on Communication, Control, and Computing*, 2003.

[100] S. Tatikonda and S. Mitter. Control under communication constraints. *IEEE Transactions on Automatic Control*, 49(7):1056–1068, July 2004.

[101] Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685, December 2009.

[102] Sebastian Thrun and Anton Schwartz. Finding structure in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS) 7*, 1995.

[103] Niko Tinbergen. *The Study of Instinct.* Clarendon PRess, 1951.

[104] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[105] Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In Vassilis Cutsuridis, Amir Hussain, and John G. Taylor, editors, *Perception-Action Cycle*, Springer Series in Cognitive and Neural Systems, pages 601–636. Springer New York, 2011.

[106] Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28):11478–83, July 2009.

[107] H. Touchette and S. Lloyd. Information-theoretic approach to the study of control systems. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):140–172, 2004.

[108] Hugo Touchette and Seth Lloyd. Information-theoretic limits of control. *Physical Review Letters*, 84(6):1156–1159, 2000.

[109] Sander G. van Dijk and Daniel Polani. Grounding subgoals in information transitions. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 105–111, Paris, France, 2011.

[110] Massimo Vergassola, Emmanuel Villermaux, and Boris I. Shraiman. 'infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409, January 2007.

[111] Satosi Watanabe. Information-theoretical aspects of inductive and deductive inference. *IBM Journal of Research and Development*, 4(2):208–231, 1960.

[112] A. C. Yu and D. Margoliash. Temporal hierarchical control of singing in birds. *Science*, 273(5283):1871–1875, September 1996.

[113] Lei Zhao, Haim H Permuter, Young-Han Kim, and Tsachy Weissman. Universal estimation of directed information. In *IEEE International Symposium on Information Theory*, pages 1433–1437, 2010.

[114] B.D. Ziebart, Andrew Maas, J.A. Bagnell, and A.K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438, 2008.