Information Integration in Embodied Systems

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation (Monografie)

von
Carlotta Langer

aus
Rheda-Wiedenbrück

2024

Date of Oral Examination:	17. May 2024
Chair of Examination Board:	Prof. Dr.-Ing Gerhard Bauch
First Examiner:	Prof. Dr. rer. nat. Nihat Ay
Second Examiner:	Prof. Dr. rer. nat. Daniel Polani
Third Examiner:	Prof. Dr. sc. nat. Verena V. Hafner

Identifiers
DOI: https://doi.org/10.15480/882.13079
ORCID: https://orcid.org/0000-0003-3536-9433

Abstract

Embodied agents inevitably live in constant interaction with their environment. In this thesis we model this interaction via the sensorimotor loop and analyze the different information flows among an agent's body, brain and environment using information theoretic methods. In particular, we highlight the interplay between the complexity of the controller of an agent and the interaction of an agent's body with its environment. The controller complexity is assessed using measures that are defined in the context of the Integrated Information Theory of consciousness. This theory aims at quantifying the amount and quality of the consciousness of a system. However, there exist various candidates for an Integrated Information measure. Hence, we compare and discuss existing information geometric measures and define two additional ones, namely the causal information integration and the ground truth Integrated Information. The former should be used if there exists an unknown exterior influence on the controller and in the case of the latter one all exterior influences are known.

The ground truth Integrated Information is then applied to an experimental setting with simplistic simulated agents. These agents are modeled by discrete, time-homogeneous Markov processes and their goal is to not touch the wall of a racetrack, their environment. We apply the information geometric *em*-algorithm to optimize the behavior of the agents. This is a known method called "Planning as Inference". Using twelve information theoretic measures we analyze the information flow among the agent's body, brain and environment thoroughly, including two measure related to the concepts Morphological Computation and Integrated Information. The former quantifies the reduction of computational cost for the controller that results from the interaction between the body and the environment. We observe that Morphological Computation and Integrated Information exhibit an antagonistic relationship. The more an agent interacts with its environment the less controller complexity is required.

While this is an intuitive result it leads to the problem that embodied intelligence is correlated with a reduced necessity of a complex controller, a brain. Here, we propose one potential solution to this problem given by the challenge of learning. In order to interact well with their surroundings the agents first have to understand the relevant dynamics of the environment. Hence, the agents have to form an internal world model that allows them to predict their next sensory state. To that end we modify the experiments and the learning algorithm to enable the agents to learn an optimal behavior as well as a good understanding of their environment.

By manipulating the accuracy of the world model we observe that the quality of the world model is related to the necessity for a complex controller. The better an agent understands its environment, the more it can interact with it, using Morphological Computation, which leads to a decrease in required controller complexity. Additionally, we observe that agents need an increased controller complexity in order to learn an accurate world model. Hence, a complex controller can facilitate a better interaction between the agent and its environment. In conclusion, the controller complexity and Morphological Computation influence each other.

Furthermore, we advance the knowledge in connection with the applied information geometric algorithms, more precisely the iterative scaling and *em*-algorithm.

Acknowledgements

First of all I would like to thank my advisor Nihat Ay for giving me the opportunity to study this interesting topic, for many helpful scientific discussions and his support during this time.

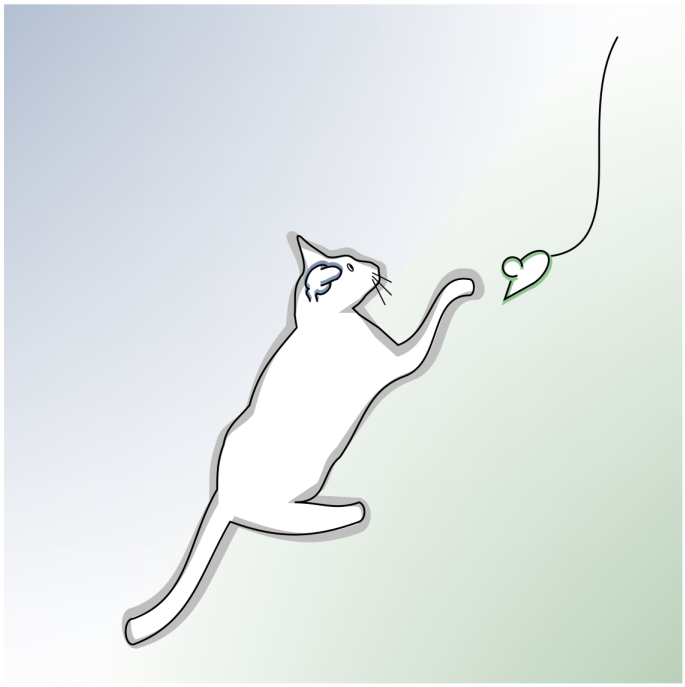
Furthermore, I am grateful for the time and the many interesting discussions with my colleagues in Hamburg, especially Csongor Várady, Jesse van Oostrum, Frank Röder, Jan Dohmen and Manfred Eppe.

I spent the first part of my time as a PhD-student at the Max-Planck Institute for Mathematics in the Sciences in Leipzig. There I would like to thank Nathaniel Virgo and Keyan Ghazi-Zahedi for their help and feedback. Additionally, I thank Bernd Sturmfels, who introduced me to algebraic geometry which led to a great collaboration with Michael Ruddy and Jane Ivy Coons.

This work was funded by the DFG within the priority program "The Active Self". I am grateful for all the stimulating discussions during the biannual meetings and the summer schools. Within this project I would like to especially thank Yasmin Kim Georgie and Verena Hafner for allowing me to gain insights into the field of robotics.

For the great support in all administrative issues I would like to thank Antje Vandenberg, Sandra Krüger, Imke Bartscher and Daniela Gascon Bosqued.

Finally, I would like to thank my family and friends, especially Chris and my siblings Franziska and Max, for their support and patience.



Contents

List of Symbols

Graphs

$An(V_1)$	Ancestral set of vertices $V_1 \subseteq V$	20
$bd(V_1)$	Boundary of a set of vertices $V_1 \subseteq V$	20
$\perp\!\!\!\perp_c$	c-separation	22
$C(G)$	Set of cliques of a graph G.	19
E	Set of edges $E = V \times V$	19
G	Graph $G = (V, E)$	19
G^m	The moral graph of G	20, 21
$pa(V_1)$	Parents of a set of vertices $V_1 \subseteq V$	20
V	Set of vertices $v \in V$	19
Y_V	Random vector corresponding to the vertices in V	19
$\mathcal{Y}_V$	State space of Y_V	19
Γ	Simplicial complex	19
$\Lambda(G)$	Set of chain components of the CG G.	21

Probability Theory

$\perp\!\!\!\perp$	Stochastic Independence	14
A_t^i	Random variable describing an actuator node	68
C_t^j	Random variable describing a controller node	33
$D_{\mathcal{Y}}(P \parallel Q)$	KL-Divergence between P and Q over the state space $\mathcal{Y}$	16
$D_{\mathcal{Y}_1\|\mathcal{Y}_2}(P \parallel Q)$	Conditional KL-Divergence between P and Q over the space $\mathcal{Y}_1 \times \mathcal{Y}_2$	16
$\mathcal{E}$	Exponential Family	17
$\mathcal{F}$	σ-algebra	14
$H_{P,Q}(Y)$	Cross entropy	16
$H_P(Y)$	Entropy	16
$H_P(Y_1\|Y_2)$	Conditional entropy	16
$I(Y_1; Y_2)$	Mutual Information	17
$I(Y_1; Y_2\|Y_3)$	Conditional Mutual Information	17
$\mathcal{M}$	Set of probability distributions	23
$\mathcal{M}_A$	Agent manifold	73
$\mathcal{M}_G$	Goal manifold	72
$MP(\mathcal{C})$	Set of stationary, discrete, time-homogeneous Markov Processes	33
n	Number of controller nodes in one point in time	33
$\tilde{P}$	Empirical distribution	23
$\hat{P}$	Probability distribution of a fully connected system	34
P, Q	Probability distributions	14
$\mathcal{P}(\mathcal{Y})$	Set of probability distributions on $\mathcal{Y}$	14
$\mathcal{P}^\circ(\mathcal{Y})$	Set of positive probability distributions on $\mathcal{Y}$	14
R	Arbitrary probability distribution	131
S_t^k	Random variable describing a sensor node	68
W	Random variable representing the exterior influence from the world	14
W_C	Random variable representing all the exterior influences on the controller nodes	35
$\mathcal{X}$	State space of X_t	14
$(X_t)_{t\in\mathbb{N}}$	Markov chain describing the agent	14
Y	Discrete random variable	14

$\mathcal{Y}$	State space of the random variable Y	14
$(\Omega, \mathcal{F}, \mathbf{P})$	Probability space	14
$\delta_{\mathcal{Y}_1}(y)$	Indicator function for the set $\mathcal{Y}_1$.	27

Measures

Φ_{CII}, $\mathcal{M}_{CII}$	Measure and set of Causal Information Integration	45
Φ_{CIS}, $\mathcal{M}_{CIS}$	Measure and set defined by CIS Integrated Information	38, 39
Φ_{EII}	Measure for the Effective Information Integration	96
Φ_G, $\mathcal{M}_G$	Measure and set of Geometric Integrated Information	37
Φ_I, $\mathcal{M}_I$	Measure and set of the Mutual Information	35
Φ_{SI}, $\mathcal{M}_{SI}$	Measure and set of Stochastic Interaction	37
Φ_T, $\mathcal{M}_T$	Measure and set of the Ground Truth Integrated Information	41
Ψ_{AE}	Measure for the Action Effect	83
Ψ_{AP}	Measure for the Actuator Prediction	109
Ψ_C	Measure for the Control	80
Ψ_{CP}	Measure for the Controller Prediction	109
Ψ_{EP}	Measure for the Environment Predictability	84
Ψ_{FP}	Measure for the Full Prediction	109
Ψ_M	Measure for the Memory	80
Ψ_{MC}	Measure for the Morphological Computation	82
Ψ_{MSI}	Measure for the Multisensory Integration	88
Ψ_R	Measure for the Reactive Control	87
Ψ_{SI}	Measure for the Sensory Information	88
Ψ_{SP}	Measure for the Sensory Prediction	83
Ψ_{Syn}	Measure for the Synergistic Information	84
Ψ_{SynP}	Measure for the Synergistic Prediction	109
Ψ_{TIF}	Measure for the Total Information Flow	77

Partition Models

$A^1 \Cap A^2$	Intersection of two partition matrices	140
$A^1 \Cup A^2$	Union of two partition matrices	140
$a^i_{j,k}$	Element in the ith block, jth row and kth column of A	136
c^ℓ_k	The kth column weight for the first ℓ partitions of A	140
$\mathcal{E}_A$	Exponential family associated with the multipartition matrix A	135
$I(a^i_j)$	Index set of the non-zero elements in the row a^i_j	140
$\mathcal{S}(i,k)$	Row index j such that $k \in I(a^i_j)$	140
$\Pi^i_j(\mathcal{Y})$	The jth set in the ith partition of the elements in the state space $\mathcal{Y}$	138

List of Abbreviations

CG Chain Graph 18, 173

CII Causal Information Integration 44

CIS Conditional Independence Statement 20

CMG Chain Mixed Graph 18, 23, 173

DAG Directed Acyclic Graph 18, 173

GRIP Generalized Running Intersection Property 135

i.i.d. independent and identically distributed 23

IIT Integrated Information Theory 30

IPS Iterative Proportional Scaling 27

MLE Maximum Likelihood Estimate 23, 135

RIP Running Intersection Property 135, 137

UG Undirected Graph 18, 173

w.r.t. with respect to 19

1 Introduction

Every agent, whether it is a human, animal, plant or robot, exists in constant interaction with its environment. It perceives information about its surrounding world, integrates that in its systems and acts accordingly. In this thesis we employ techniques from information theory in order to analyze the impact of the different processes among the body, brain and environment of an agent. We describe the theoretical framework of this thesis in more detail in Section 1.2, after an introduction to the intuition behind this approach.

1.1 Motivation

Figure 1.1 depicts a sketch of an agent interacting with its environment. This agent senses information through its eyes, processes this information in its brain, the controller, and acts on the world, including the balloon, using its hands. All these interactions among body, brain and environment are closely linked and have thereby an influence on each other. The properties of an agent's body have an impact on the necessary complexity of the controller, while in turn the capabilities of the controller restrict the interactions of body and environment. We emphasize these ideas with the following example.

First, we consider cats and their extraordinary ability to survive high falls. In [Vnuk04; Papazoglou01] the authors discuss data of cats falling from different floors in high buildings. In these studies they consider at least the second floor and over 90% of the cats survived. One cat described in [Diamond88] sustained only minor injuries from a fall from the 32th floor to concrete. There are different aspects of the cats morphology that lead to this high survival rate. Cats have an exceptionally well developed vestibular system, which is the sensory system for the sense of balance, and a righting reflex. One reason that greatly improves their ability to safely perform small jumps is a flexible spine, which allows to adjust the back muscle stiffness to absorb kinetic energy. This was analyzed in more detail in [Zhang14a; Zhang14b] using a pressure pad and cats jumping from up to 1.8m.

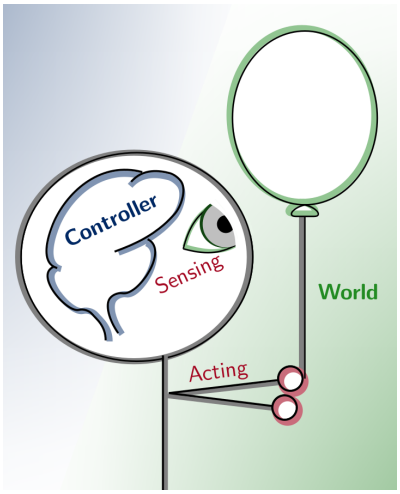


Figure 1.1: An agent interacting with its environment.

The human morphology does not equip us with such an advantage, instead we have to carefully learn different techniques if we want to jump safely down from a high point. Our brains therefore have to compensate for a morphology that is less adapted to jumping compared to felines. The necessary complexity of the processes in our brain for performing the task is higher compared to the cat.

This is in line with Brooks' approach to intelligence, published in [Brooks91a]:

> "It is soon apparent, when 'reasoning' is stripped away as the prime component of a robot's intellect, that the dynamics of the interaction of the robot and its environment are primary determinants of the structure of its intelligence."

On the other hand, athletes trained in different landing strategies in sports such as parkour or parachuting can employ various strategies to reduce the impact force of the landing on their joints, as analyzed in [Puddle13], and therefore minimize the risk for injuries, discussed in [Aziz20; Kwok03]. An experienced athlete does not have to consciously place every joint in the correct position. The way the body reacts in this situation and the necessary movement succession has already been learned and is known to the control

architecture. Therefore, a well-trained controller facilitates a better interaction between the body and the environment. Hence, these two different processes are tightly coupled.

Our objective is to analyze this relationship by observing the information flows in an embodied agent under different morphological circumstances. In order to thoroughly analyze the different flows we create an experiment with simplistic agents. We are in particular interested in the complexity of the control architecture, measured by an Integrated Information measure, as described in more detail in Section 1.2.3 and Chapter 3.

1.2 Theoretical Framework

This section briefly introduces the main components of the theoretical framework applied in this thesis.

1.2.1 The Sensorimotor Loop

The interaction of an agent with its environment, as described in the previous section, can be modeled by the sensorimotor loop, also called perception-action loop or action-perception cycle in for example [Klyubin07; Klyubin04; Salge14; Tishby11]. A sketch of this is depicted in Figure 1.2.

It illustrates that the actions generated by an agent's actuators influence the environment and therefore alter the information the sensors perceive about the world. Here the sensors and actuators produce the corresponding internal signals and are therefore part of the cognitive system. This information can get processed in the controller or additionally evoke a direct stimuli-response from the actuators. The controller sends commands to the actuators.

The idea of this feedback loop between the environment and the agent can also be found in Uexküll's function-circle [Uexküll92] or, with a more detailed control structure, in the work by Kirsh [Kirsh94].

In [Powers73] the author postulates in Chapter 4 that

> "feedback, when correctly analyzed, is the central and determining factor in all observed behavior."

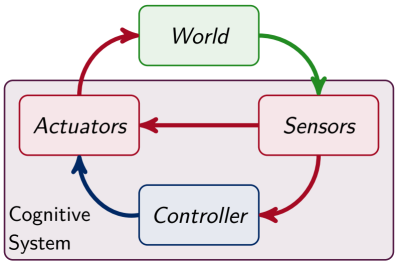


Figure 1.2: Sketch of the Sensorimotor loop.

Furthermore, the sensorimotor loop is also applied in a branch of cognitive science called embodied cognition. There the proponents theorize that the sensory data is structured by the interaction of the agent with its environment and that therefore this interaction is crucial for cognition, as discussed in for example [Bourgine04; Varela91; Pfeifer07; Stewart10]. Overviews over the landscape of embodied cognition approaches can be found in, for instance, [Wilson02; Kiverstein09] and we further discuss the importance of embodiment in Section 1.2.4.

The interaction between the agent and its environment can be modeled by dynamic systems, as described in [Thelen94; Der12]. However, in this thesis the different parts of the sensorimotor loop are associated with discrete random variables and the interactions among them correspond to discrete conditional probability distributions. More precisely, we model the system as a Bayesian network, similar to the treatment in [Ay15b; Klyubin05; Klyubin04]. In [Ay14] the authors analyze the sensorimotor loop from a causal perspective.

We refer to the interactions among the components of the sensorimotor loop as information flows, discussed in more detail in the next section.

1.2.2 The Information Flows

One focus of this thesis is to analyze the information flow in artificial agents under different circumstances. Here we understand the notion of an information flow from the information theoretic perspective. In [Shannon48] Shannon introduces a mathematical theory of communication, which constitutes the foundation of information theory. There he quantifies the capacity of a communication channel and the possible compression rate of a message, which lead to the concepts of mutual information and entropy, introduced in Section 2.2. Here we apply such information theoretic measures to quantify the influence one component in the sensorimotor loop has on another. In [Polani07; Tishby11] the authors discuss this setting and conclude that information theory is well suited to describe the mechanisms in the sensorimotor loop. This approach is also applied in, for instance, [Bialek01; Klyubin04; Polani09].

Further examples include [Klyubin07], where the authors maximize the information flow through the whole system as a learning objective. The concepts of information and entropy are applied in [Touchette04] to define conditions under which a system is perfectly controllable or observable. Emphasizing the importance of the sensory input entropy and mutual information are utilized in [Sporns04; Lungarella05] to examine how an agent actively structures its sensory and motor data. Moreover, in [Lungarella06] the authors find that the information flow in the agent can be affected by the morphology of an agent's body. In [Ay21] the author addresses problems with interpreting the mutual information causally, arising from its symmetric properties.

Analyzing the information flow can be approached in two different ways. First, one can quantify the properties of the conditional distributions that describe the mechanisms of the system. This is a purely theoretical method. A second option is to actively perturb the system and observe the impact on the information flows. In Pearl's causality theory, introduced in for example [Pearl88; Pearl09], interventions are formalized by the *do*-operation, which fixes the value of a part of the system and thereby models an experimental intervention, as described in [Pearl19]. In [Ay08b] the authors use this theory to define information flows in causal systems. One caveat of this approach is that direct interventions change the system and therefore might not accurately reflect the mechanisms in the original situation.

In this thesis we do not use direct interventions, but we instead examine the information flow theoretically. Moreover, we introduce different types of architectures of the agents to be able to observe the impact of certain structures more closely. These additional architectures inhibit a part of the information flow. This allows us to thoroughly examine all the various information flows among the different elements of the agent and the world.

One element in the sensorimotor loop, that we are particularly interested in, is the internal complexity of the control architecture. We quantify this by calculating the amount of Integrated Information, as discussed in the next section.

1.2.3 Complexity and Integrated Information

The complexity of a system can be apparent for an observer but hard to grasp conceptually. A complex system should neither be organized in a regular structure nor completely random but something in between. There exist various approaches to complexity, discussed for example in [Kinsner10; Ladyman13; Gell-Mann99]. In this thesis we call a system complex if the whole is more than the sum of its parts.

Following this notion of complexity we use Integrated Information measures for the analysis of the controller. These measures quantify the difference between the whole system and one in which the different parts of the controller do not interact with each other. Hence, if these measures are zero, then we are able to model this system by a split system without any functional difference. The question of how to define a split system leads to various

candidate measures that we discuss in Chapter 3. There we define different measures including one, in the case of only known influences on the controller, that we apply to the simulated agents in Chapter 4 and 5.

These measures are a part of the broader "Integrated Information Theory of Consciousness" developed mainly by Tononi and colleagues [Tononi94; Tononi08; Oizumi14]. The proponents of this theory aim at quantifying the quality and amount of consciousness of a system by the way it integrates information. We discuss the different branches of this theory and their history in Section 3.1. The Integrated Information measures are mostly restricted to the controller of the agent, nonetheless, some studies also relate Integrated Information to the behavior and environment.

The authors of [Edlund11] measure the Integrated Information in simulated, evolving agents in a maze and conclude that Integrated Information increases along with the fitness of the agents. Similarly, the authors of [Albantakis14] come to the result that the Integrated Information of an agent needs to increase for the agent to be able to navigate in a more complex environment. Furthermore, in [Albantakis15] the authors observe that a high Integrated Information value increases the likelihood of a rich, dynamical behavior. Focusing on the internal structure of evolving agents and how they represent their environment leads the authors of [Albantakis18] to the conclusion that an important aspect of an agent is not solely its behavior but also its inner workings. There they measure the internal information integration of an agent as well as the joint entropy of the world and the controller given the sensory state. The latter measures how much the brain represents about the agent's environment and is discussed by the authors of [Marstaller13]. They concentrate on the role of representation and observe that their measure increases with the lifetime of the agents.

In this thesis we aim to analyze the complexity of the controller explicitly in relation to the remaining information flows among body, brain and environment of an agent. One important concept regarding the interaction of the body and environment is called Morphological Computation.

1.2.4 Embodiment and Morphological Computation

The field of embodied artificial intelligence highlights the importance of considering an embodied agent with a control architecture embedded in the sensorimotor loop. This fundamentally differs from the traditional approach to artificial intelligence, which typically describes a computational process independent of its environment. These classical systems in general do not perform well in real-world situations, as addressed in [Pfeifer03]. Hence, the field of embodied artificial intelligence aims at understanding natural systems by studying situated agents.

In, for example, [Polani07; Pfeifer06a] the authors argue that intelligence requires a body. Moreover, the author of [Brooks91a] lists the following reasons for considering embodied agents.

> "First, only an embodied intelligent agent is fully validated as one that can deal with the real world. Second, only through a physical grounding can any internal symbolic or other system find a place to bottom out, and give 'meaning' to the processing going on within the system."

Furthermore, also higher level cognitive concepts such as the self are thought to be connected to an agents body. The author of [Gal00] defines the concept of a "minimal self" with two key components, namely the sense of agency and ownership. Especially the latter, the sense that I am the one having an experience, is tightly linked to body representations and therefore embodied agents. Works towards modeling a minimal self utilize real robots, as discussed in [Georgie19; Nguyen21; Hafner20].

In this thesis we also follow the embodied approach and aim to analyze the influence the body has on the brain and vice versa. To that end we quantify both, the interaction of the body with the environment and the information flows inside the controller respectively, using the same framework. This enables us to relate them to each other.

The interaction between body and environment leads to a concept called "Morphological Computation". This describes the reduction of computational complexity for the controller that results from the interaction between the body and the world, as defined in, for example, [Ghazi-Zahedi19]. There exist various definitions of the notion of Morphological Computation and also different ways in which the body can lift the burden of the brain, as discussed in, for example, [Müller17]. In the field of soft robotics, for instance, the qualities of the flexible tissue of the robots enable the generation of movements via relatively simple control architectures. An example with an octopus-inspired robotic arm can be found in [Nakajima13a]. We address these aspects of Morphological Computation further in Section 4.1.1.

In Section 3 in [Pfeifer09] the authors characterize the relationship between Morphological Computation and the control architecture in the following way.

> "Second, there is a kind of trade-off or balance: the better the exploitation of the dynamics, the simpler the control, the less neural processing will be required."

One consequence of this relationship is the "cheap design principle", formulated in the book [Pfeifer06a]. It states that an agent's body should be constructed with the goal to exploit the properties of their environment as best as possible. The authors of [Montúfar15] formalize the notion of cheap control architectures, meaning ones that exploit the agent's embodiment, in the context of universal approximation.

Hence, a control architecture does in general not need to be sophisticated in order to create complex behavior, as demonstrated in [Braitenberg84] by Braitenberg's famous thought experiment. There he introduces multiple agents with very simplistic sensor-actuator couplings that lead to surprisingly complex behavioral patterns.

One additional insight, pointed out by Braitenberg, is that it is tempting to use terminology attributed to reasoning, consciousness or even emotion when we observe and describe the behavior of simulated agents. In a psychological study that Heider and Simmel publish in [Heider44] test subjects tend to describe the movement of geometrical objects in terms of living beings. This study and later replications are discussed in [Kück06]. Therefore, we want to point out that in this thesis we neither attribute consciousness nor reasoning to the simulated agents. They are merely a minimal example used to analyze the interactions among the different parts of the sensorimotor loop.

The main results of our experiments are outlined in the next section.

1.3 Main Results and Outline

Here we outline the structure of this thesis and its main outcomes.

There are two different types of results presented in this thesis, namely experimental and theoretical. We observe experimental results generated with simulated agents. In order to define the theoretical framework of these experiments thoroughly we need to further advance the theoretical knowledge about this setting. Hence, we prove mathematical results in the context of complexity measures for Integrated Information and additionally analyze two information geometric algorithms, namely the iterative scaling and *em*-algorithm. We highlight the mathematical results in the summary below by marking them by the corresponding theorem or proposition number.

The following chapter, Chapter 2, provides an introduction to the mathematical background. This includes a brief definition of discrete probability distributions and some

selected information theoretic concepts. Furthermore, we introduce the Boltzmann Machine that we utilize in the following chapters to generate examples. Afterwards, we define exponential families and graphical models leading to the notions of conditional independence and causality. Lastly, Section 2.5 consists of the discussion of two information geometric algorithms. These algorithms calculate the maximum likelihood estimation of an initial distribution in different situations and are called the *em*-algorithm and the iterative scaling algorithm.

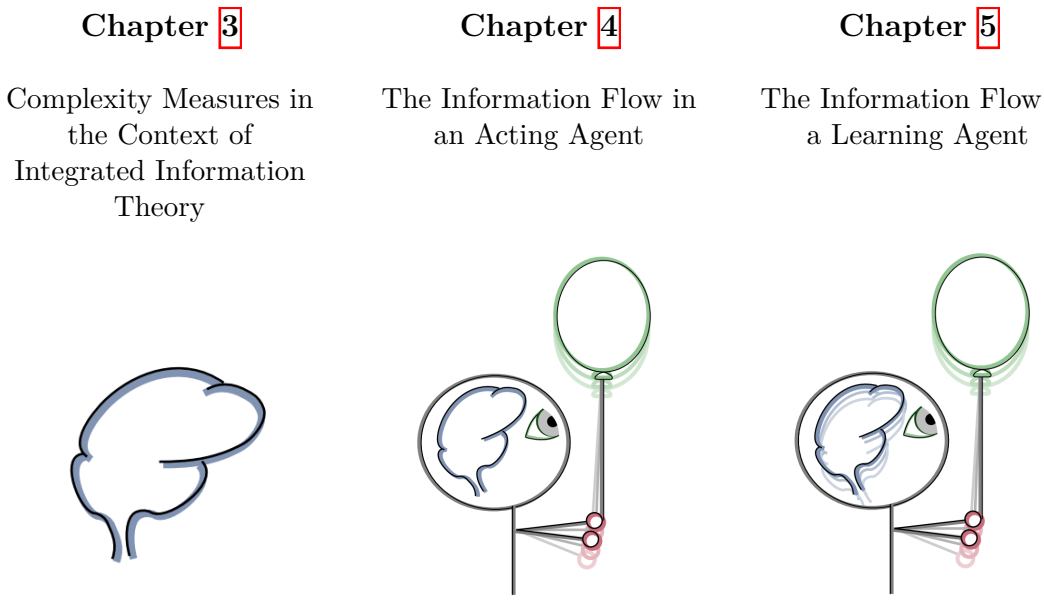


Figure 1.3: Visualization of the three main chapters of this thesis.

The three main parts of this thesis are visualized in the Figure 1.3. We start by discussing different Integrated Information measures in Chapter 3. These aim at quantifying the complexity of the controller of an agent and, hence, this chapter is visualized via a stylized brain in Figure 1.3. We apply one of these measures to acting agents in Chapter 4 in order to observe the interplay between the complexity of the controller and the interaction of the agent with the world. This setting is symbolized by an agent interacting with its environment, given by a balloon. In this case the agent only acts in its environment and does not learn from the interaction. The analysis of the learning process is the crucial next step discussed in Chapter 5. Therefore we depict the brain of the agent in Figure 1.3 as changing, adapting to its environment.

We now discuss the chapters and their results in more detail.

Results of Chapter 3

In Chapter 3 we focus on finding an appropriate measure for the complexity of the controller. The discussed complexity measures can be seen in the context of the Integrated Information Theory. Hence, we start that chapter with an introduction to the Integrated Information Theory of consciousness, which aims at quantifying the amount and quality of the consciousness of a system. Depending on the version of the theory and the type of system it is applied to there exist various Integrated Information measures, as we discuss in more detail in Section 3.1.

We contribute to this discussion in the context of information geometric measures, compared in, for example, [Oizumi16b; Kanwal17; Mediano18]. The core idea of these

measures is to quantify the difference between the original system and one without the ability to integrate information, called the split system. This difference is calculated by the Kullback-Leibler-divergence, called KL-divergence. The discussed measures differ in the way they define the split system.

The authors of [Oizumi16b; Amari18] postulate two properties a valid Integrated Information measure should fulfill. As mentioned in Section 1.2.3, the main goal of these measures is to assess the importance of the connections among different controller nodes, across different points in time. Therefore, Postulate 1 aims at ensuring the removal of these connections, which we call "causal cross-connections". They guarantee the lack of the causal cross-connections by defining Markov conditions that a valid measure should satisfy.

Secondly, they also postulate that any Integrated Information measure should be bounded from above by the mutual information, which removes all the information flows that originate from the controller nodes in time point t and point to nodes in $t+1$. In that case an undirected connection among the controller nodes at time point $t+1$ is used to account for an exterior influence.

The validity of Postulate 2 is debated in [Kanwal17] because always accounting for the undirected connection among the nodes at time point $t+1$, even when there might be no exterior influence, can lead to an incorrectly low value for Integrated Information. In that case the undirected connection can compensate for a part of the causal cross-connections. This discussion leads us to the conclusion that we need to distinguish between three different situations, depicted in Figure 1.4. These cases differ in the amount of exterior influence on our system. Either there is no influence or the exterior influence can be known or unknown.

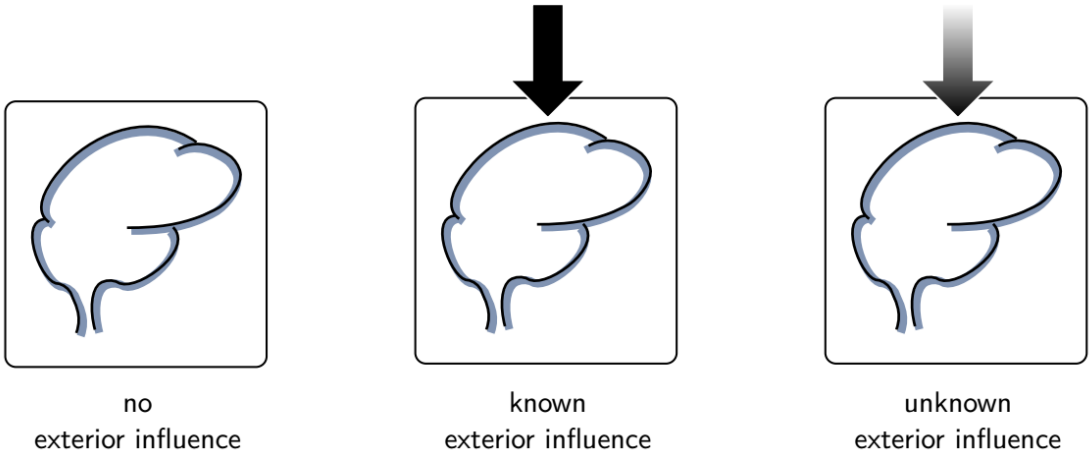


Figure 1.4: Visualization of the three different ways the system can be influenced from outside.

In the case of no exterior influence there exists a measure called "stochastic interaction", published in [Ay01; Ay15a], that quantifies only the strength of the causal cross-connections, as desired. The question of finding the corresponding measures for a known and unknown exterior influence leads us to the definition of two new measures, namely a measure defined by the ground truth in Section 3.3 and the causal information integration in Section 3.4.

The measure defined by the ground truth explicitly includes a known exterior influence in its system. It assesses the casual cross-connections and thereby satisfies Postulate 1. This measure is not bounded by the mutual information because it operates on the larger state space, which includes the exterior influence. Its properties, such as a closed form solution as the sum of conditional mutual information terms and that it can be expressed via conditional independence statements, are shown in the Propositions 3, 4 and 5. Since

this measure only quantifies the causal-cross connections, we are able to draw the following conclusion:

> *In the case of entirely known exterior influences on the system the ground truth defines a measure that leads to the correct way of calculating the Integrated Information in our setting.*

In the case of an unknown exterior influence we are not able to access the ground truth. Instead we define a new measure in Section 3.4, where the exterior influence is given by a latent variable. Since this satisfies the Postulates 1 and 2 and the corresponding split system has a geometrical interpretation, which allows for a causal interpretation of this model, we name this measure causal information integration. If this measure is zero, then the system can be explained as a limit of marginalized distributions without causal cross-connections, as shown in Proposition 6. Hence, we conclude the following:

> *If there exists an unknown exterior influence on the system, then the causal information integration measure should be applied.*

For each size of the latent space the measure can then be calculated using the *em*-algorithm, as specified in Section 3.4.1 including Proposition 7. We argue in favor of only using latent spaces with sizes that are appropriate in the case of the discussed system. If the exterior influence on a system is multiple magnitudes larger than the system itself, then the internal information flow might not be relevant for the observed behavior of the system.

Furthermore, we are able to show that it is not possible to define a different measure that also satisfies Postulate 1 by adding further external influences to the system.

> "Theorem 3.4.2": *It is not possible to gain a truly larger system that satisfies Postulate 1 by adding further hidden influences to the model of causal information integration.*

The authors of [Oizumi16b; Amari18] suggest an additional measure that also satisfies both postulates. There the split system is defined using conditional independence statements (CIS). Hence, we name this measure CIS Integrated Information. The split system of this measure does not have a graphical representation and is therefore harder to interpret causally. It is a lower bound on the causal information integration, however, numerical evaluations suggest that these two measures are in general not equal.

In Section 3.5 we summarize the relationships between the different measures and discuss examples demonstrating them. In particular, we show in Proposition 8 that the ground truth Integrated Information is an upper bound of the causal information integration.

Results of Chapter 4

One of the frameworks in which one can use the ground truth Integrated Information is discussed in Chapter 4. There we present experiments in which simple simulated agents are faced with the task of not touching the walls of their environment, a racetrack.

The main focus of this chapter is to analyze the dynamics of the different information flows inside the agents under changing morphological circumstances. The assumption that the morphology of an agent's body and its interaction with the environment has a significant impact on an agent's brain is a key element in the field of embodied cognition, as discussed in, for example, [Varela91; Bourgine04; Pfeifer07; Stewart10; Pfeifer06a]. In this thesis we develop a framework in which we are able to quantify all the information flows among the different parts of an agent and its environment. Thereby we can directly observe the

interplay between the various influences which allows us to experimentally verify previously made intuitive hypotheses. Before we highlight our results we first describe the experiments in more detail.

An agent in our experiments is modeled via a discrete, time-homogeneous Markov-Process. It is equipped with two binary sensors to detect the walls and can perform four movements, fast forward, slow forward and forward to the left or right. In this experiment the agent does not move directly inside the environment, but we instead sample the relationship between an agent's sensory state, its movement and its next sensory state, beforehand. The resulting conditional probability distribution is called empirical world model. This model reflects the dynamics of the world relevant for the agent, as shown in Proposition 9.

> "Proposition 9:" *Under the assumption that the world has only an impact on the sensors, even when considered over time, the empirical world model results from marginalizing over the world states and therefore captures the dynamics of the environment of an agent.*

This approach allows us to calculate the best strategies for the agents, purely on a theoretical basis. The applied method is called "Planning as Inference", introduced in for example [Attias03; Toussaint06; Toussaint09], and here we apply the information geometric *em*-algorithm in order to find the best strategies for the agents. More precisely, we iterate to the conditional distributions that belong to strategies closest to one that guarantees success, as defined in Section 4.3.1. It is important to note that the agents have to perform two steps of their movement before the information flow in the controller becomes relevant for the behavior of the agents.

Additionally, we consider three different types of agents, discussed in Section 4.4. Firstly, there are the "fully coupled" agents with unrestricted connections among the controller, actuator and sensor nodes. Secondly, we introduce "controller driven" agents. These are not able to directly send information from the sensors to the actuators. Therefore, they are incapable of direct, reactive actions. Finally, in the case of the "reactive control" agents the controller has no influence on the actuators and therefore the behavior of the agents. These different types of agents allow us to observe the impact a missing connection has practically, not only theoretically.

For every type of agents we take 100 random input distributions and use the *em*-algorithm to calculate the strategies with the best likelihood of not touching a wall in the next two points in time. We then apply twelve different information theoretic measures to the resulting information flows. These measures are introduced in Section 4.5 and each of them is the result of minimizing the KL-divergence between the original system and the set of split ones in which a specific part of the information flow does not exist. This missing part is the one that we want to quantify.

The resulting measures can be divided into three categories, namely the information flows coming from the controller, the flow inside the empirical world model and the sensory information inside the agent, discussed in the Sections 4.5.2, 4.5.3 and 4.5.4. Additionally, the measure called "total information flow" is an upper bound for all the other measures because in that case all the interactions among the sensor, actuator and controller nodes are removed in the split system.

By applying all these measures we are able to observe their dynamics in relation to each other. One of these is a measure for Morphological Computation. It quantifies the reduction of computational cost for the controller that results from the interaction between the body and environment of an agent, as discussed in more detail in Section 4.1.1. This definition leads to the natural assumption that there exists a trade-off between Morphological Computation and the complexity of the control architecture. The more an agent interacts

with its environment and exploits the dynamics, the simpler can the control architecture be constructed, as described in [Pfeifer06a; Pfeifer09]. One remaining problem with this theory is pointed out by the authors of [Pfeifer06b] in Section 5:

> "One problem with the concept of morphological computation is that while intuitively plausible, it has to date defied serious quantification efforts: We would like to be able to ask 'How much computation is actually being done?'"

Hence, there exist various measures for Morphological Computation, candidates are discussed in for example [Klyubin05; Ghazi-Zahedi13; Ghazi-Zahedi17a]. In [Ghazi-Zahedi19] the author compares various measures and addresses their shortcomings and advantages. In this thesis we mainly focus on one of the measures that was deemed to be suitable in [Ghazi-Zahedi19] and that is consistent within our framework. Thereby we are able to quantify Morphological Computation as well as the complexity of the control architecture in the same framework. This allows us to directly observe the dynamics of these measures and test the above mentioned trade-off hypothesis. The complexity of the controller is measured by an Integrated Information measure, discussed in Chapter 3, that we define using the ground truth.

We influence how much the agent perceives about its environment by manipulating the reach of the agent's sensors. This has a direct impact on the agent's ability to successfully interact with its environment.

The results of the importance of the interaction between the body and the environment, namely Morphological Computation, and the controller complexity, given by Integrated Information, are depicted in Figure 1.5. We observe an antagonistic relationship between these two and that leads us to the following conclusion:

> *The more an agent relies on interacting with the environment, using Morphological Computation, the less complex the controller of an agent needs to be.*

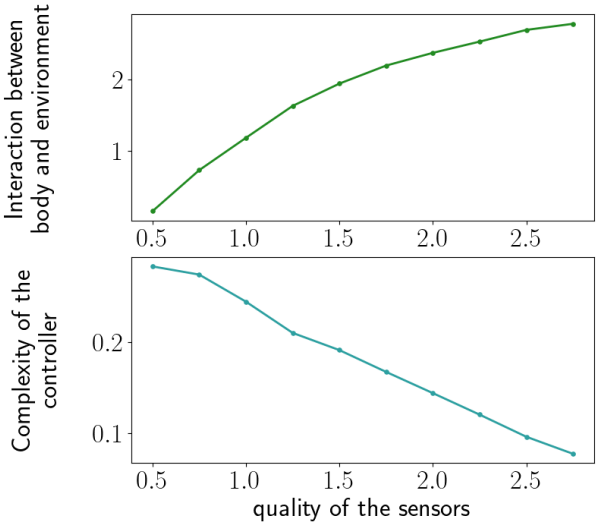


Figure 1.5: Results of the measures Morphological Computation (top) and Integrated Information (bottom) presented in Section 4.6

So, in the case of a low quality of the sensors the agents are not able to interact well with their environment and they have a high controller complexity. This confirms the intuition that there exists a trade-off between Morphological Computation and the controller complexity.

Integrated Information measures are solely defined on the controller system and therefore this measure does not consider the behavior of an agent at all. However, in this thesis we also want to understand the impact of the complexity of the controller on the actions of an agent. We observe that the high Integrated Information value in case of low quality sensors has next to no impact on the behavior of the agent. In that case the importance of a direct reaction is high and the impact of the controller nodes on the actuators is very low. Hence, the importance of the information flow in the controller also depends on the information flowing from it to the actuators.

Additionally, the information sent from the sensors to the controller should make an impact, such that the Integrated Information is meaningful. Therefore, we define an indicator for these three information flows as the product of all of them and call it the effective information integration. A sketch of the information flows considered by this quantity is given in Figure 1.6.

The impact of the Integrated Information in the controller on the behavior of an embodied agent depends additionally on the information flowing to and from the controller nodes. Therefore it is not sufficient to only calculate an Integrated Information measure but we need to consider the effective information integration.

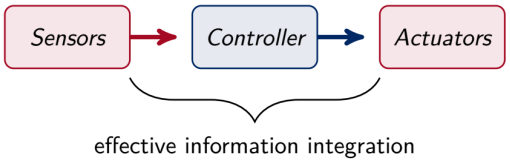


Figure 1.6: Sketch of the information flows that the effective information integration considers.

The observed antagonistic relationship between Morphological Computation and Integrated Information lead to the question of why a well-adapted system would need a complex controller, and with that Integrated Information, at all. One possible answer is given by the challenge of learning, which we address in the next chapter.

Results of Chapter 5

Chapter 5 adapts and extends the framework from Chapter 4 in order to analyze the information flow in agents while these agents interact with their environment and learn. The goal of the agents remains not to touch the walls, but if they do touch them they are stuck as long as one of the sensors detects a wall. Hence, they need to turn on the spot so that they face away from a wall and then they are able to move away. Furthermore, the Markov process describing the agents is changed in a way that allows a more direct impact of the controller values on the behavior of the agent.

In order to be able to optimize their behavior w.r.t. the goal of not touching the walls the agents need to predict their next sensory state. One option of gaining such a prediction is to sample the sensor and actuator values in connection with the next sensory state. The resulting distribution is called an "empirical world model". Agents with access to this model are called "ideal" because they have an optimal understanding of their environment.

On the other hand, agents normally don't have a direct access to the dynamics of the world and therefore need to form an internal world model that predicts the next sensory state given the actuator and controller states. These agents have to learn their internal world models. Therefore we adapt the *em*-algorithm to incorporate two goals, as defined in Section 5.4.

The modified em-algorithm optimizes the internal world model and the behavior of the agent simultaneously. There we iterate between optimizing the strategy w.r.t. the goal and increasing the accuracy of the internal world model.

This internal world model adds to the complexity of the controller, hence, in order to determine the controller complexity we not only calculate the Integrated Information value but also a complexity measure for the internal world model, called synergistic prediction as defined in Section 5.5. Additionally, we consider six of the measures defined in Chapter 4 and calculate three further measures on the information flow in the internal world model.

In Section 5.6 we first discuss the dynamics of these measures for the ideal agents. There the results suggest that a complex controller might not be necessary for agents with access to their empirical world model. In that case the information flow from the controller to the actuators is close to zero after the first few steps. This suggests that an accurate understanding of the environment might lead to a better interaction with it and thereby to a lower controller involvement.

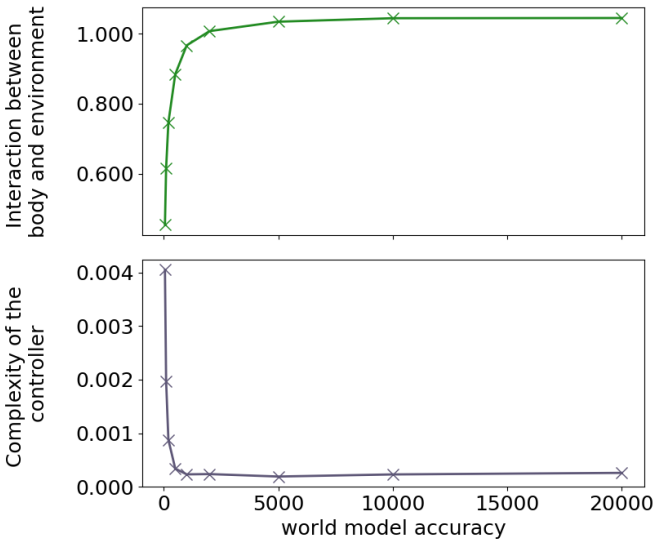


Figure 1.7: Results of the Morphological Computation (top) and the effective information integration (bottom) from Section 5.6.2

We further analyze this by considering agents that are able to update their empirical world model for only the first steps. Thereby, we change the accuracy of the world model. The results are depicted in Figure 1.7 where the accuracy of the world model increases with the number on the x-axis. We observe that an agent with a good understanding of its environment, meaning it has an accurate world model, has a higher Morphological Computation and lower controller complexity compared to agents with an inaccurate world model. This leads to the following conclusion:

> *The better an agent understands its environment, the more it can exploit the interactions between body and environment and the less controller complexity is needed.*

Next we consider the agents that have to form an internal world model. The ones that are reasonably successful in avoiding the walls have a high controller complexity in the first few steps which then monotonically decreases, as depicted in Figure 1.8. This leads us to the hypothesis that a high Integrated Information might be necessary in order to learn an accurate internal world model. When this world model is learned it can be used to improve the interaction with the environment, which leads to a reduction of the complexity of the controller. We analyze this further by considering "split" agents.

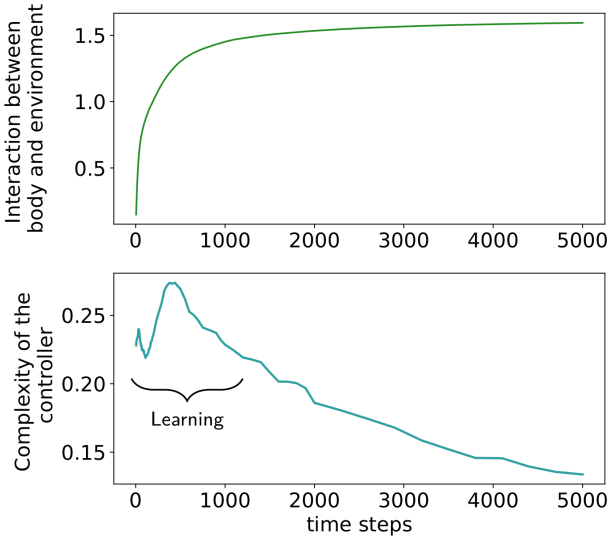


Figure 1.8: Results of the Morphological Computation (top) and Integrated Information and synergistic prediction (bottom) presented in Section 5.6.3

These are not able to integrate information in their controller with the result that they perform significantly worse. A synergistic measure applied to the world model additionally shows that the few split agents that succeed in the task have to combine the information from the controller and the actuators directly in the world model. In conclusion, this leads to the following result:

> *An increased controller complexity is necessary to learn an accurate world model. The agent needs to combine different information sources, which can happen either between the controller nodes, in form of Integrated Information, or inside the world model, measured by the synergistic prediction.*

A summary and discussion of the results of Chapter 5 is given in Section 5.7.

Results of Chapter 6 and 7

The Chapter 6 includes further theoretical results about the two information geometric algorithms, namely the *em*-algorithm and the iterative scaling algorithm, that we apply in this thesis.

In Section 6.1 we discuss the *em*-algorithm in case of a model that includes a latent variable. Amari et al. show in [Amari92] that the global minimum in the visible space is equal to the global minimum in the larger space that includes latent variables. We apply this algorithm in Chapter 3 in the context of the causal information integration. However, in that context also the size of the state space of the latent variables is unknown. Hence, we analyze ways to increase the size of this state space in a plausible way. More precisely, we define three methods in which we use the local minimum in a smaller latent space to gain an initial distribution for the *em*-algorithm in the larger one. Here we prove the following result.

> "Proposition 10": *The "safe" method allows us to increase the size of the latent space and guarantees that the local minima do not get worse, compared to the smaller space.*

The last section, Section 6.2, takes a closer look at the iterative scaling algorithm. This algorithm is used in this thesis to calculate the geometric Integrated Information in Chapter 3, the synergistic information measure in Chapter 4 and the synergistic prediction in Chapter 5. It works by iteratively projecting to linear sets in order to find their intersection which is the MLE.

Even small examples show that the number of necessary projections can vary greatly with the chosen representation of these sets. For a subclass of hierarchical models, called decomposable hierarchical models, there exists a property that guarantees that the iterative scaling algorithm converges to the MLE in one cycle. This is called the running intersection property, was defined in [Haberman74] and analyzed in more detail in [Vomlel99]. We prove a property for partition models, which are a generalization of hierarchical models, that also ensures one-cycle convergence of the iterative scaling algorithm. This property is named GRIP, generalized running intersection property.

> "Theorem 6.2.2": *The iterative scaling algorithm converges for partition models with a matrix representation that satisfies the GRIP in one cycle to the MLE.*

In Chapter 7, we draw conclusions and discuss possible directions for further research that result from this thesis. In particular, we summarize the insights from Chapter 4 and Chapter 5 regarding the relationship between Morphological Computation and Integrated Information in Section 7.2.1.

2 Mathematical Background

In this thesis we analyze the information flow in a discrete system from an information theoretic perspective. Therefore this chapter provides a short introduction into the mathematical foundations, such as information theory, graphical models and information geometric algorithms. Hence, there are no new results in this chapter, but it should increase the clarity of the thesis.

We model the information flow among different parts of a system using the language of discrete probability theory. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with the σ-algebra $\mathcal{F}$ on the set Ω and the probability measure $\mathbf{P}$. The random variable Y is defined as a function

$$Y : \Omega \to \mathfrak{Y}$$

with the countable set $\mathfrak{Y}$, such that for all $y \in Y(\Omega)$

$$\{\omega \in \Omega | Y(\omega) = y\} \in \mathcal{F}.$$

The range of Y is called the state space $\mathcal{Y} \subseteq \mathfrak{Y}$. The probability distribution of that random variable is denoted by $P(Y)$ and defined as

$$P(y) = \mathbf{P}(\{\omega \in \Omega : Y(\omega) = y\})$$

for $y \in \mathcal{Y}$. Probability distributions will be denoted by P, Q.

Let $Y = (Y_1, Y_2)$ be a random vector with the state space $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$. Then the marginal distribution on $\mathcal{Y}_1$ can be calculated by

$$P(y_1) = \sum_{y_2 \in \mathcal{Y}_2} P(y_1, y_2)$$

for all $y_1 \in \mathcal{Y}_1$. To simplify the notation we write

$$\sum_{y} \text{ instead of } \sum_{y \in \mathcal{Y}}.$$

The set of probability distributions on $\mathcal{Y}$ is denoted by $\mathcal{P}(\mathcal{Y})$ and the set of strictly positive distributions on $\mathcal{Y}$ by $\mathcal{P}^\circ(\mathcal{Y})$.

The random variables Y_1 and Y_2 are called conditionally independent given Y_3, if the probability distribution of Y_3 is positive and

$$P(y_1, y_2 | y_3) = P(y_1 | y_3) P(y_2 | y_3)$$

holds for all $(y_1, y_2, y_3) \in \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}_3$. Then we write $Y_1 \perp\!\!\!\perp Y_2 | Y_3$. Properties of conditional independence statements are analyzed in, for example, [Dawid79; Studený89] and we discuss them more in Section 2.4. Further details on probability theory can be found in for example [Brémaud17] or [Athreya06].

The systems analyzed in this thesis are modeled as discrete Markov chains $(X_t)_{t \in \mathbb{N}}$, where $\mathcal{X}$ is the state space of the random vector $X_t = (X_t^1, \ldots, X_t^\nu)$, $\nu \in \mathbb{N}$. A Markov chain is a discrete stochastic process for which the Markovian condition

$$P(X_{t+1} | X_t, \ldots, X_0) = P(X_{t+1} | X_t)$$

holds. We also refer to it as Markov process. Additionally, the Markov chains discussed in this thesis are time-homogeneous, meaning that the transition probability does not depend on the timestep

$$P(X_{t+1} | X_t) = P(X_{t+2} | X_{t+1})$$

for all $t \in \mathbb{N}$.

The Markov chain $(X_t)_{t \in \mathbb{N}}$ describes the behavior of the agent throughout this thesis and an additional influence from the environment will be denoted by W.

2.1 Boltzmann Machine

The type of Markov chain that we use to gain example distributions is related to a Boltzmann machine. This is a fully connected recurrent neural network that is defined in the following way.

Let us consider ν variables $X_t = (X_t^1, \dots, X_t^\nu)$ with the binary state space $\mathcal{X} = \{-1, 1\}^\nu$. The transition probability can then be written as

$$P(x_{t+1}^j | x_t) = \frac{1}{1 + e^{-2\beta \sum_{i=1}^{\nu} v_{ij} x_t^i x_{t+1}^j}}, \tag{2.1}$$

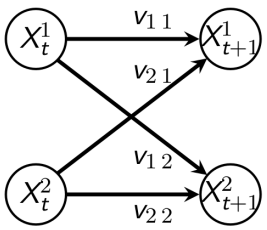


Figure 2.1: The weights corresponding to the connections for $\nu = 2$.

where the weights v_{ij} of the connection from X_t^i to X_{t+1}^j are stored in the matrix $V \in \mathbb{R}^{\nu \times \nu}$. The connections from X_t to X_{t+1} with the corresponding weights are displayed in Figure 2.1.

Note that this figure is not a graphical model corresponding to the stationary distribution, as will be defined in Section 2.4, but merely displays the connections of the conditional distribution of $X_{t+1}^j = x_{t+1}^j$ given $X_t = x_t$. The connections in the Boltzmann machine are assumed to be undirected, such that the connection matrix V is symmetric.

The inverse temperature $\beta > 0$ regulates the coupling strength of the connections from X_t to X_{t+1}. For β close to zero the different variables are almost independent and as β increases the connections become stronger. The function in the exponent of the denominator in (2.1)

$$E(V, X_t, X_{t+1}) = \sum_{i=1}^{\nu} v_{ij} x_t^i x_{t+1}^j$$

is called the energy function. The units in this model behave stochastically like the spins in an Ising model, which is an array of atoms that can take only the states $\pm$ 1 and was introduced by Ising in [Ising25]. Hence, this model is also known as a weighted Ising model or binary auto-logistic model as described in Reference [Winkler03] Example 3.2.3.

A Boltzmann machine can also be seen as a variation of a Hopfield network, as described by Hinton and Sejnowski in [Hinton83]. The Hopfield network, introduced in [Hopfield82], consists of binary threshold nodes and has therefore a deterministic update rule. Hence, the Boltzmann machine can also be described as a stochastic Hopfield network, as discussed in Section 43.1 of [MacKay03].

The updates of the states in a Boltzmann machine are typically done asynchronously. Instead, here we follow the approach described in [Kanwal17] and use these networks with fixed weight matrices V and a synchronous update in order to generate a stationary distribution depending on β. A distribution is called stationary if applying the transition probability does not change it. Since the transition probability is positive, there always exists a unique stationary distribution, see for instance Theorem 4.3.1. in [Winkler03]. For more details regarding the connections among Ising model, Hopfield networks or Boltzmann machines refer to, for example, [Peretto84; MacKay03; Hertz91].

We are calculating the stationary distribution $\bar{P}$ by starting with a random, positive initial distribution P^0 and then multiply by (2.1) in the following way

$$P^{t+1}(x_t) = \sum_{x_t} P^t(x_t) \prod_{j=1}^{n} P(x_{t+1}^j | x_t),$$

for all $(x_t, x_{t+1}) \in \mathcal{X} \times \mathcal{X}$.

This leads to $\bar{P} = \lim_{t\to\infty} P^t$. Note that the result of this synchronous update, $\bar{P}$, is in general not equivalent to the outcome of the Boltzmann machine with an asynchronous update rule. Another difference between the classical Boltzmann machine and our application is that we do not restrict V to be symmetric.

We use tools from information theory, described in the next section, in order to analyze properties of the resulting stationary distribution $\bar{P}$.

2.2 Information Theory

In 1948 C.E. Shannon published in [Shannon48] an approach to a general theory of communication. This influential work includes the quantification of the uncertainty of the outcome of a random variable, called entropy, as lower bound of compression.

Definition 1 (Entropy)**.** Let Y be a random variable on $\mathcal{Y}$. Then the entropy $H_P(Y)$ is defined as

$$H_P(Y) = -\sum_y P(y)\, \log P(y)$$

with the convention that $0 \cdot \log(0) = 0$. For the random vector $Y = (Y_1, Y_2)$ on $\mathcal{Y}_1 \times \mathcal{Y}_2$ the conditional entropy $H_P(Y_1|Y_2)$ is given by

$$H_P(Y_1|Y_2) = -\sum_{y_1,y_2} P(y_1, y_2)\, \log P(y_1|y_2).$$

Let P, Q be probability distributions on $\mathcal{Y}$ then the cross entropy $H_{P,Q}(Y)$ is calculated as

$$H_{P,Q}(Y) = -\sum_y P(y)\, \log Q(y).$$

The Kullback-Leibler-divergence, also called KL-divergence or relative entropy, quantifies the difference between two probability distributions.

Definition 2 (KL-Divergence)**.** Let $Y = (Y_1, Y_2)$ be a random vector on $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$. The KL-divergence and the conditional KL-divergence are defined as

$$D_{\mathcal{Y}}(P \parallel Q) = \sum_y P(y)\, \log\left(\frac{P(y)}{Q(y)}\right) = H_{P,Q}(Y) - H_P(Y)$$

$$D_{\mathcal{Y}_1|\mathcal{Y}_2}(P \parallel Q) = \sum_y P(y)\, \log\left(\frac{P(y_1|y_2)}{Q(y_1|y_2)}\right)$$

with the conventions that $0 \cdot \log \frac{0}{0} = 0$, $0 \cdot \log \frac{0}{Q(y)} = 0$ and $P(y) \cdot \log \frac{P(y)}{0} = \infty$ for $P(y) > 0$.

This measures how much the uncertainty of the random variable increases if we use Q instead of P. More details can be found in for example [Cover06]. Here we include the following properties:

1. $D_{\mathcal{Y}}(P \parallel Q) \geq 0$
2. $D_{\mathcal{Y}}(P \parallel Q) = 0$ if and only if $P = Q$

Proofs of these properties can be found in [Cover06] in Theorem 2.6.3. The first property above implies

$$H_{P,Q}(Y) \geq H_P(Y). \tag{2.2}$$

Definition 3 (Mutual Information)**.** Let (Y_1, Y_2) be a random vector on $\mathcal{Y}_1 \times \mathcal{Y}_2$ with the distribution P. Then the mutual information is the KL-divergence between $P(Y_1, Y_2)$ and the product distribution $P(Y_1)P(Y_2)$

$$I(Y_1; Y_2) = \sum_{y_1} \sum_{y_2} P(y_1, y_2) \log \left(\frac{P(y_1, y_2)}{P(y_1)P(y_2)} \right).$$

Now, let (Y_1, Y_2, Y_3) be a random vector on $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}_3$ with the distribution P and $P(y_3) > 0$. The conditional mutual information of the random variables Y_1 and Y_2 given Y_3 is defined as $I(Y_1; Y_2|Y_3)$

$$\begin{aligned} I(Y_1; Y_2|Y_3) &= \sum_{y_1} \sum_{y_2} \sum_{y_3} P(y_1, y_2, y_3) \log \left(\frac{P(y_1, y_2|y_3)}{P(y_1|y_3)P(y_2|y_3)} \right) \\ &= \sum_{y} P(y) \log \left(\frac{P(y_1|y_2, y_3)}{P(y_1|y_3)} \right). \end{aligned}$$

Using property 2 of the KL-divergence we see that $I(Y_1; Y_2|Y_3) = 0$ if and only if the equality $P(y_1|y_2, y_3) = P(y_1|y_3)$ holds for all $y \in \mathcal{Y}$, which can be expressed as a conditional independence statement $Y_1 \perp\!\!\!\perp Y_2|Y_3$.

2.3 Exponential Families

This section is a short introduction to a widely used class of statistical models, called exponential families. A statistical model is a set of probability distributions that can be parameterized explicitly by a set of parameters Θ or defined implicitly by constraints, for example conditional independence statements. We introduce the latter in the next section. A longer discussion of exponential families can be found in for example [Ay17; Amari16; Geiger98; Geiger01].

Definition 4 (Exponential Family)**.** Let F be a vector space spanned by a set of functions

$$\{f_\ell | f_\ell : \mathcal{Y} \to \mathbb{R}, \ell \in \{0, \dots, d-1\}\}.$$

The exponential family defined by F is given by

$$\mathcal{E}(F) = \left\{ P \in \mathcal{P}(Y) | P(y) = \frac{1}{Z(\theta)} e^{\sum_\ell \theta_\ell f_\ell(y)}, \theta_\ell \in \mathbb{R} \right\},$$

where $\theta = (\theta_0, \dots, \theta_{d-1})$ is the vector consisting of the parameters of this model. This way of parameterizing an exponential family is called natural or canonical. The function

$$Z(\theta) = \sum_y e^{\sum_\ell \theta_\ell f_\ell(y)}$$

is the normalizer, also called partition function. Because of the normalization two elements of F that differ by an additive constant lead to the same distribution P. In order to assure uniqueness we assume that the function that maps every $y \in \mathcal{Y}$ to 1 is in F.

The family of distributions above is also called linear exponential family or log-affine model because taking the logarithm leads to

$$\log P(y) = \sum_\ell \theta_\ell f_\ell(y) - \log(Z(\theta)).$$

The linearity of the space F allows to define F via a matrix. This alternative definition can be found in for example in [Drton09a], Section 6.2 of [Sullivant18] and Section 6.2.1 of this thesis.

An important class of subfamilies of exponential families are curved exponential families.

Definition 5 (Curved Exponential Family)**.** A curved exponential family is a subfamily of an exponential family in which there exists a mapping from a parameter space $\Theta' \to \Theta$, $\epsilon \mapsto \theta(\epsilon)$ such that the image is a d'-dimensional smooth manifold in $\mathbb{R}^d$

$$\mathcal{E}(F, \epsilon) = \left\{ P \in \mathcal{P}(Y) | P(y) = \frac{1}{Z(\theta)} e^{\sum \theta(\epsilon) f(y)} \right\}.$$

In order to define the second important class of subfamilies we need the following concepts. A partition of M is a set of nonempty, disjoint subsets of M whose union is M. A stratification of a subset $M \subseteq \mathbb{R}^d$ is a finite partition $\{M_1, \ldots, M_k\}$, $k \in \mathbb{N}$, of M such that

(1) each stratum $M' \in \{M_1, \ldots, M_k\}$ is a d_k-dimensional smooth manifold in $\mathbb{R}^d$

(2) and if $M' \cap \overline{M}'' \neq \emptyset$, then $M' \subseteq \overline{M}''$ and and $dim(M') < dim(M'')$ for $M', M'' \in \{M_1, \ldots, M_k\}$.

A stratified set is a set that has a stratification. The dimension of a stratified set is the largest dimension of a stratum.

Definition 6 (Stratified Exponential Family)**.** A stratified exponential family is a subfamily of an exponential family in which there exists a mapping from a parameter space $\Theta' \to \Theta$, $\epsilon \mapsto \theta(\epsilon)$ such that the image is a d'-dimensional stratified set in $\mathbb{R}^d$.

These two kinds of subfamilies, curved and stratified exponential families, play an important role in the context of graphical models, discussed in the next section.

2.4 Graphical Models, Conditional Independence and Causality

In this section we introduce the concepts of graphical models. A graphical model is a statistical model in which the structure of the probability distributions can be expressed by a graph. These graphs are an intuitive way to display conditional independence statements, which are called Markov properties in this context.

We discuss four different types of graphs in the following sections, namely undirected graphs (UGs), directed acyclic graphs (DAGs), chain graphs (CGs) and chain mixed graphs (CMGs). Figure 2.2 displays an example for each of these graphs. A more detailed introduction to graphical models can be found in [Lauritzen96].

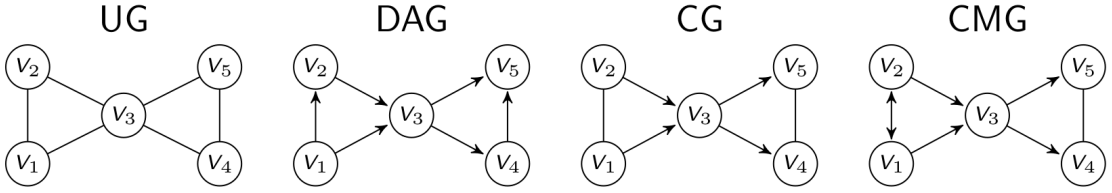


Figure 2.2: Examples of an Undirected Graph (UG), a Directed Acyclic Graph (DAG), a Chain Graph (CG) and a Chain Mixed Graph (CMG).

Additionally, we discuss the concept of causality in the context of DAGs. The interplay between graphical models corresponding to DAGs and their encoded independence statements provides the foundation of Judea Pearl's causality theory.

2.4.1 Undirected Graphs

A graph G is a pair $G = (V, E)$ with vertex set V and edge set E. An undirected edge consists of an unordered pair of elements in V, $e = \{v_1, v_2\} \in E$ with $v_1, v_2 \in V$. If all edges are undirected then the graph is called an undirected graph. A clique of a graph G is a subset of vertices, $A \subseteq V$, in which each pair of distinct elements is connected by an edge $e \in E$. If a subgraph has this property then it is also called complete. Let $C(G)$ be the set of cliques of G. A path is a sequence of distinct vertices $v_1 = v^1, v^2, \dots, v^d = v_2$ such that $(v^\ell, v^{\ell+1}) \in E$ for $\ell \in \{1, \dots, d-1\}$. For two vertices v_1, v_2 a subset $V_3 \subseteq V$ is called a (v_1, v_2)-separator if every path from v_1 to v_2 intersects V_3 and V_3 separates two sets $V_1, V_2 \subseteq V$ if it is an (v_1, v_2)-separator for every pair $v_1 \in V_1$ and $v_2 \in V_2$.

Now we associate the vertices of the graph with random variables in order to connect a graph with a set of probability distributions. Let $\mathcal{Y}_V$ be the state space of the random vector $Y_V = (Y_{v_1}, Y_{v_2}, \dots)$ corresponding to the vertices in V. We denote a state of the full random vector by y_V and an element of the state space $\mathcal{Y}_{v_1}$ of only $v_1 \in V$ by y_{v_1}.

In the case of an undirected graph the graphical model is the exponential family in which the spanning functions of the vector space $F(G)$ in Definition 4 are defined by the cliques of G. Then $f \in F(G)$ if there exist real-valued functions ϕ_A that only depend on the configurations of the cliques $A \in C(G)$ and parameters θ_A such that

$$f(y_V) = \sum_{A \in C(G)} \theta_A \phi_A(y_A),$$

for all $y_V \in \mathcal{Y}_V$.

Therefore, there exist non-negative real-valued functions h_A such that every distribution in the graphical model factors according to the cliques of the graph

$$\begin{aligned} P(y_V) &= \frac{1}{Z(\theta)} e^{\sum_{A \in C(G)} \theta_A \phi_A(y_A)} \\ &= \prod_{A \in C(G)} h_A(y_A), \end{aligned}$$

for all $y_V \in \mathcal{Y}_V$.

A positive probability distribution factorizes according to an undirected graph G if and only if it satisfies the global Markov property for undirected graphs, as defined below. This is known as Hammerlsey-Clifford theorem, see for example Theorem 2.9 in [Ay17].

Definition 7 (Global Markov Property for UGs)**.** A distribution $P \in \mathcal{P}^\circ(\mathcal{Y}_V)$ satisfies the global Markov property relative to the UG G if for any triple $V_1, V_2, V_3 \subseteq V$ of disjoint sets such that V_3 separates V_1 and V_2 in G the conditional independence statement $Y_{V_1} \perp\!\!\!\perp Y_{V_2} | Y_{V_3}$ holds for P.

A generalization of graphical models are hierarchical models. Here the interaction structure is not given by the cliques of a graph but by a simplicial complex. Let Γ be a subset of the powerset $2^{|V|}$. Then we call Γ a simplicial complex if for every $\gamma \in \Gamma$ and $\gamma' \subseteq \gamma$ already $\gamma' \in \Gamma$ follows. The elements of a simplicial complex are the faces of Γ and all inclusion maximal faces are called facets. In this case the spanning functions of the space $F(G)$ are then defined with respect to (w.r.t.) the facets of Γ. For instance, the set of spanning functions can be chosen such that for each facet there is an indicator function for one state of the variables in this facet. This approach corresponds to the matrix representation in Example 4. More details on hierarchical models can be found in Section 2.9 of [Ay17] and in Section 6.2.1 of this work.

In Section 6.2 we introduce a further generalization of hierarchical models that we call partition models, also described in [Coons24].

2.4.2 Directed Graphs and Causality

Now we discuss graphical models that are defined in relation to a graph with only directed edges, visualized by arrows, and the property that there exist no directed cycles. A directed edge is an ordered pair of elements in V. These graphs are then called directed acyclic graphs (DAGs).

Let $v \in V$ be a vertex belonging to the DAG G. Then we call $pa(v_1) \subseteq V$ the set of all parents of v_1, meaning the set of all vertices $v_2 \in V$ such there is a directed edge from v_2 to v_1. In this case v_1 is called a child of v_2. A distribution P factors according to G if we are able to write

$$P(y_V) = \prod_{v \in V} P(y_v | y_{pa(v)}),$$

for all $y_V \in \mathcal{Y}_V$.

In the case of DAGs the corresponding graphical model is not an exponential family anymore but a curved exponential family, see for example [Geiger01; Geiger98] for more details.

Similarly to the undirected case we have a global Markov property for directed graphs. In order to define this property we first need to additionally introduce the concepts of an ancestral set and a moral graph.

The boundary $bd(V_1)$ of a set $V_1 \subseteq V$ is the set of vertices in $V \setminus V_1$ that are parents or neighbors to vertices in V_1. Two vertices are called neighbors if they are connected by an undirected edge. Hence, in the case of DAGs the boundary of V_1 only consists of the parents of the vertices in V_1 denoted by $pa(V_1)$. If $bd(v_1) \subseteq V_1$ for all $v_1 \in V_1$ we call V_1 an ancestral set. The intersection of ancestral sets is again an ancestral set and therefore there exists a smallest ancestral set for any $V_1 \subseteq V$. This smallest ancestral set of V_1 is denoted by $An(V_1)$.

The moral graph of G is an undirected graph denoted by G^m that consists of the same vertex set as G and in which two vertices v_1, v_2 are connected if either they were already connected by an edge in G or if they have a common child. Now we are able to define the global Markov property for DAGs.

Definition 8 (Global Markov Property for DAGs)**.** Let $P \in \mathcal{P}^\circ(\mathcal{Y}_V)$ be a distribution on $\mathcal{Y}$ and G a DAG. P satisfies the global Markov property with respect to G if for any triple (V_1, V_2, V_3) of disjoint subsets of V such that V_3 separates V_1 from V_2 in $(G_{An(V_1 \cup V_2 \cup V_3)})^m$, the moral graph of the smallest ancestral set containing the union $V_1 \cup V_2 \cup V_3$, the conditional independence statement $Y_{V_1} \perp\!\!\!\perp Y_{V_2} \mid Y_{V_3}$ holds for P.

Note that alternatively one can define the global Markov property via a new separation criterion on DAGs called d-separation, see Proposition 3.25 in [Lauritzen96].

Analogous to the case of the UGs we also have the equivalence between the distributions satisfying the global Markov property for a DAG and the ones factorization according to this DAG, see Theorem 3.27 in [Lauritzen96].

These global Markov properties allow us to visualize, via the graphical representation, exactly the probability distributions that satisfy a set of Conditional Independence Statements (CISs). In [Dawid79] the author analyzes conditional independence statements in general and defines some of their properties. Pearl calls four of these properties the conditional independence axioms and any set of triplets that satisfies these are then called semigraphoids. A detailed analysis of probabilistic independence statements can be found in [Studený05] and a discussion about semigraphoids is given in [Studený97].

Graphical models generated by DAGs are the key element in Pearl's causality theory. Inspired by human inferential reasoning he proposes in [Pearl85a] to use these models, that he terms "Bayesian networks", to simulate causal mechanisms. The idea to use DAGs as causal diagrams and the underlying independence structure as causal independence developed into an influential theory of causality, described in for example [Pearl85b; Pearl88; Pearl09], with the interplay between CISs and their graphical representation as its foundation. In Section 1.3 of [Pearl09] the author emphasizes the importance of a graphical representation with the following statement:

> "It seems that if conditional independence judgments are by-products of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal Bayesian networks."

Bayesian networks are broadly accepted as a model to represent probabilistic knowledge and are used for a wide range of tasks such as prediction, inference or learning the causal structure of raw data, see for example [Darwiche09; Ben-Gal08; Borgelt09].

2.4.3 Chain Graphs

In this thesis, we are mainly using chain graphs (CGs). These are graphs that consist of directed and undirected edges such that we are able to partition the vertex set into subsets $V = \Lambda_1 \cup \cdots \cup \Lambda_m$, called chain components. The chain components have the properties that all edges between different subsets are directed, all edges between vertices of the same chain component are undirected and that there are no directed cycles between chain components. Therefore UGs are CGs with only one chain component and DAGs can be seen as CGs, in which every chain component includes only one vertex. Hence the following factorization according to a CG can be seen as a combination of the factorizations in the directed and undirected case.

Let $\Lambda(G)$ be the set of chain components of G. A distribution factorizes with respect to a chain graph G if the distribution can be written as follows

$$P(y_V) = \prod_{\lambda \in \Lambda(G)} P(y_\lambda | y_{pa(\lambda)}),$$

for all $y_V \in \mathcal{Y}$ and where the structure of $P(y_\lambda | y_{pa(\lambda)})$ can be described in more detail. Let $\lambda_\star$ be the undirected graph with the vertex set $\lambda \cup pa(\lambda)$ and the edges are the ones between elements in $\lambda \cup pa(\lambda)$ that exist in G and additionally all edges between elements in $pa(\lambda)$, even if these elements are not connected in G. Additionally, let $A(\lambda), \lambda \in \Lambda$ be the set of all subsets of $\lambda \cup pa(\lambda)$ that are complete in a graph $\lambda_\star$. Then there are non-negative functions ϕ_a such that

$$P(y_\lambda | y_{pa(\lambda)}) = \prod_{a \in A(\lambda)} \phi_a(y_a).$$

If λ consists of only one vertex, then $\lambda_\star$ is already complete. Chain graphs without hidden variables are curved exponential families [Geiger01].

We now define the global Markov property for chain graphs, defined in for example [Frydenberg90]. A discussion of different types of Markov properties for chain graphs is given in [Drton09b]. Now we need to broaden the definition of a moral graph for CGs. The moral graph of a CG G is an undirected graph denoted by G^m that consists of the same vertex set as G. Two vertices v_1, v_2 are connected in G^m if and only if either they were

already connected by an edge in G or if there are vertices v_3, v_4 belonging to the same chain component such that $v_1 \to v_3$ and $v_2 \to v_4$. In addition to the parents of a vertex we here connect the parents of the same chain component.

Definition 9 (Global Markov Property for CGs)**.** Let $P \in \mathcal{P}^\circ(\mathcal{Y}_V)$ and G a CG. P satisfies the global chain Markov property with respect to G if for any triple (V_1, V_2, V_3) of disjoint subsets of V such that V_3 separates V_1 from V_2 in $(G_{An(V_1 \cup V_2 \cup V_3)})^m$ the CIS $Y_{V_1} \perp\!\!\!\perp Y_{V_2} \mid Y_{V_3}$ holds for P.

Since we are only considering positive discrete distributions, we also have the equivalence between the global Markov property for CGs and the factorization, see Theorem 4.1 from Reference [Frydenberg90] combined with the Hammersley-Clifford theorem, for example, Theorem 2.9 in Reference [Ay17].

Although the set of distributions that factor according to a graph and the set that satisfies the global Markov property are identical the actual graph associated with these sets does not need to be unique. This leads to the concept of Markov equivalence. Two graphs are called Markov equivalent if they have the same associated Markov properties. In the case of UGs two graphs are only equivalent if they are identical but for CGs this is not true in general, see for example [Andersson97]. We are able to determine whether a chain graph is Markov equivalent to its underlying UG graph, in which all directed edges are turned into undirected ones, with the following criterion from [Frydenberg90]. A chain graph has the same Markov properties as its underlying undirected graph if and only if the boundary of every chain component is complete. Further results on the Markov equivalence of CGs and DAGs can be found in [Andersson97].

2.4.4 Chain Mixed Graphs

Now we discuss chain mixed graphs (CMG) that arise when we marginalize over vertices of a chain graph. A CMG has in addition to directed and undirected edges also bidirected edges, called arcs, and no semi-directed cycles. Two vertices connected by an arc are called spouses and those connections arise when they have a common influence that we have marginalized over. The probability distributions associated with these graphs are then stratified exponential families, as shown in [Geiger01], and we do not have any factoriziation according to these graphs, but we are able to define the conditional independence relations on these graphs using the following c-separation criterion, defined for example in [Sadeghi16] in Section 4.

There we use the concepts of a walk and a collider section. A walk is a list of vertices $v^1, \ldots, v^d$, $d \in \mathbb{N}$, such there is an edge or arrow from v_ℓ to $v_{\ell+1}$, $\ell \in \{1, \ldots, \ell - 1\}$. A set of vertices connected by undirected edges is called a section. If there exists a walk including a section such that an arrow points at the first and last vertices of the section $\to v_1 - \cdots - v_2 \leftarrow$ then this is called a collider section.

Definition 10 (c-separation)**.** Let V_1, V_2 and V_3 be disjoint sets of vertices of a graph. A walk π is called a c-connecting walk given V_3 if every collider section of π has a node in V_3 and all non-collider sections are disjoint. The nodes V_1 and V_2 are called c-separated given V_3 if there are no c-connecting walks between them given V_3 and we write $V_1 \perp\!\!\!\perp_c V_2 | V_3$.

The following algorithm from [Sadeghi16] converts a chain graph with latent variables into a chain mixed graph with the conditional independence structure of the marginalized chain graph.

Definition 11 (Marginalization for Chain Mixed Graphs)**.** Let W be the set of vertices over which we want to marginalize. The following algorithm produces a chain mixed graph (CMG) with the conditional independence structure of the marginalized chain graph.

1. Generate an edge between v_1 and v_2 as in Table 2.1, steps 8 and 9, if the edge of the same type does not already exist.
2. Generate an edge as in Table 2.1, steps 1 to 7, between the endpoints of every tripath with inner node in W if the edge of the same type does not already exist. Apply this step until no other edge can be generated.
3. Remove all nodes in W.

1	$v_1 \leftarrow w \leftarrow v_2$	generates	$v_1 \leftarrow v_2$
2	$v_1 \leftarrow w - v_2$	generates	$v_1 \leftarrow v_2$
3	$v_1 \leftrightarrow w - v_2$	generates	$v_1 \leftrightarrow v_2$
4	$v_1 \leftarrow w \rightarrow v_2$	generates	$v_1 \leftrightarrow v_2$
5	$v_1 \leftarrow w \leftrightarrow v_2$	generates	$v_1 \leftrightarrow v_2$
6	$v_1 - w \leftarrow v_2$	generates	$v_1 \leftarrow v_2$
7	$v_1 - w - v_2$	generates	$v_1 - v_2$
8	$w \rightarrow v_1 - \cdots - v_3 \leftarrow v_2$	generates	$v_1 \leftarrow v_2$
9	$w \rightarrow v_1 - \cdots - v_3 \leftrightarrow v_2$	generates	$v_1 \leftrightarrow v_2$

Table 2.1: Types of edge induced by marginalizing over the variable w.

In the following sections graphs are only used in relation to a graphical model and therefore we use the names of the random variables directly as names for the vertices and shorten Y_V to Y.

2.5 Information Geometric Algorithms to calculate the Maximum Likelihood Estimation

Let $\mathcal{M}$ be a set of probability distributions. Given a set of observations $\{d_1, d_2, \dots\}$ we want to select the distribution $P \in \mathcal{M}$ to which the data fits best. This means that we want to maximize the probability of the data. Assuming that the observations are independent and identically distributed (i.i.d.) with $d_i \in \mathcal{Y}$ maximizing the likelihood w.r.t. $\mathcal{M}$ leads to

$$\max_{P \in \mathcal{M}} \prod_{d_i} P(d_i).$$

Every data point d_i corresponds to a state $y \in \mathcal{Y}$, such that $d_i = y$. Let $u(y)$ be the number of times d_i appears in the set of observations. Then this maximization is equivalent to maximizing the log-likelihood

$$\max_{P \in \mathcal{M}} \log \left(\prod_y P(y)^{u(y)} \right) = \max_{P \in \mathcal{M}} \sum_y u(y) \, \log \left(P(y) \right)$$

because of the monotonicity of the logarithm.

Definition 12 (Maximum Likelihood Estimation)**.** Given a set of probability distributions $\mathcal{M}$ on $\mathcal{Y}$ and a random variable Y with the empirical distribution $\tilde{P}$. Then the Maximum Likelihood Estimate (MLE) is given by

$$P^\star = \arg\max_{P \in \mathcal{M}} \sum_y \tilde{P}(y) \, \log P(y).$$

This leads to the same result as minimizing the KL-divergence with respect to the second argument, since

$$\begin{aligned}\arg\max_{P\in\mathcal{M}} \sum_y \tilde{P}(y)\, \log P(y) &= \arg\min_{P\in\mathcal{M}} \left(\sum_y \tilde{P}(y)\, \log \tilde{P}(y) - \sum_y \tilde{P}(y)\, \log P(y) \right) \\ &= \arg\min_{P\in\mathcal{M}} D_{\mathcal{Y}}(\tilde{P} \parallel P).\end{aligned} \tag{2.3}$$

In Information Geometry this is called performing an m-projection to $\mathcal{M}$, as we discuss in the next section. Hence, this allows us to use information geometric algorithms to calculate the MLE. Next we discuss two different algorithms, the *em*-algorithm and the iterative scaling algorithm, that can be applied depending on the nature of the set $\mathcal{M}$.

2.5.1 The em-Algorithm

The *em*-algorithm is a well known information geometric method to minimize the KL-divergence between two sets. It was proposed by Csiszár and Tusnády in 1984 in [Csiszár84] as alternating minimization algorithm and its usage in the context of neural networks with latent variables was described for example by Amari et al. in [Amari92]. In this case the *em*-algorithm is applied to find the MLE of the observed data, as we discuss in the next section in more detail. The expectation-maximization EM-algorithm, described in [Dempster77], used in statistics is equivalent to the *em*-algorithm in many cases, including the ones discussed in this thesis, as described in more detail in [Amari95] or in Section 5.3 of [Csiszár04].

In order to discuss the steps of the *em*-algorithm we first define two different projections based on the KL-divergence.

Definition 13 (e- and m-projection)**.** Let $\mathcal{M}$ be a set of probability distributions on the state space $\mathcal{Y}$. We perform an Information projection, or e-projection, of a distribution P^0 to $\mathcal{M}$ by minimizing the KL-divergence with respect to the first argument. Minimizing with respect to the second argument is called a reverse Information projection or m-projection.

$$\begin{aligned} e\text{-projection} \quad & \inf_{P\in\mathcal{M}} D_{\mathcal{Y}}(P \parallel P_0) \\ m\text{-projection} \quad & \inf_{P\in\mathcal{M}} D_{\mathcal{Y}}(P_0 \parallel P) \end{aligned}$$

The names, e- and m-projection, result from the relationship to exponential and mixture connections, as described in more detail in the Sections 2.4 and 2.7.2. of [Ay17] or [Amari85].

The *em*-algorithm works by iteratively projecting between two sets of probability distributions that we call $\mathcal{M}_1$ and $\mathcal{M}_2$. The projection to $\mathcal{M}_1$ is given by an e-projection and the projection to $\mathcal{M}_2$ is an m-projection. A sketch of this iterative process is depicted in Figure 2.3.

Let $Q^0 \in \mathcal{M}_2$ be an arbitrary initial distribution. Then we project this to $\mathcal{M}_1$ via an e-projection

$$P^0 = \arg\inf_{P\in\mathcal{M}_1} D_{\mathcal{Y}}(P \parallel Q_0).$$

Then we perform an m-projection to $\mathcal{M}_2$

$$Q^1 = \arg\inf_{Q\in\mathcal{M}_2} D_{\mathcal{Y}}(P_0 \parallel Q).$$

Repeating this leads to

$$P^{\ell} = \arg\inf_{P \in \mathcal{M}_1} D_{\mathcal{Y}}(P \parallel Q^{\ell}),$$
$$Q^{\ell+1} = \arg\inf_{Q \in \mathcal{M}_2} D_{\mathcal{Y}}(P^{\ell} \parallel Q).$$

The convergence of this algorithm is given by the following result.

Proposition 1 (Theorem 8 from Reference [Amari92])**.** *Performing the em-algorithm as defined above leads to the monotonically decreasing relations*

$$D_{\mathcal{Y}}(P^{\ell} \parallel Q^{\ell}) \geq D_{\mathcal{Y}}(P^{\ell+1} \parallel Q^{\ell}) \geq D_{\mathcal{Y}}(P^{\ell+1} \parallel Q^{\ell+1})$$

and equality holds only for the fixed points $(P', Q') \in \mathcal{M}_1 \times \mathcal{M}_2$ *of the projections*

$$P' = \arg\inf_{P \in \mathcal{M}_1} D_{\mathcal{Y}}(P \parallel Q')$$
$$Q' = \arg\inf_{Q \in \mathcal{M}_2} D_{\mathcal{Y}}(P' \parallel Q).$$

Proof of Proposition 1. This follows immediately from the definitions of the e- and m-projections. □

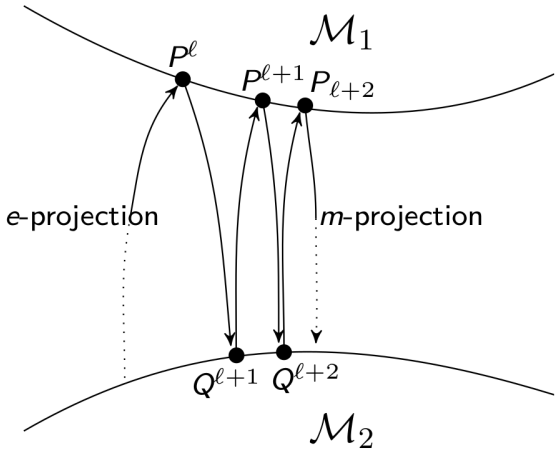


Figure 2.3: Sketch of the *em*-algorithm.

If the two sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are convex, compact and if additionally the starting distribution $Q^0 \in \mathcal{M}_2$ is strictly positive, then this algorithm converges to the global optimum as shown in Corollary 5.1 of [Csiszár04]. However, this is not the case in most applications including the ones discussed in the following chapters. Hence, in these settings the *em*-algorithm is guaranteed to converge but it might converge towards a local minimum, as discussed in, for example, [Amari95] or in Section 5.3 of [Csiszár04]. Therefore we apply the *em*-algorithm in each case multiple times with random input distributions, which leads in most cases to multiple local minima. Then we are able to take the minimum of the outcomes of the different runs as an estimate of the global minimum. Additionally, we take a closer look at these minima in Example 1.

As discussed in [Sullivant18] the behavior of the *em*-algorithm is not well understood yet. In [Kubjas15] the authors characterize the set of fixed points for two visible, discrete random variables from an algebraic geometric perspective.

The *em*-Algorithm in Case of Latent Variables

The *em*-algorithm can be used to find the MLE of partially observed data. In order to do that we distinguish between the known and latent part of the random vector y and write $y = (y_k, y_h) \in \mathcal{Y} = \mathcal{Y}_k \times \mathcal{Y}_h$. Here the h denotes the latent variables Y_h and stands for "hidden". Then we are able to define $\mathcal{M}_1$ as the set of all those probability distributions whose marginals on the known variables equal the empirical distribution of the observed data

$$\mathcal{M}_1 = \left\{ P \in \mathcal{P}(\mathcal{Y}) | P(y_k) = \tilde{P}(y_k), \forall y_k \in \mathcal{Y}_k \right\}. \tag{2.4}$$

Analogously, the second set, $\mathcal{M}_2$, is also defined on the known and the hidden states $\mathcal{M}_2 \subseteq \mathcal{P}(\mathcal{Y}_k \times \mathcal{Y}_h)$. In the following chapters $\mathcal{M}_2$ is given by a graphical model. Marginalizing to the known states leads in case of the elements in $\mathcal{M}_1$ to only one distribution $\tilde{P}(Y_k)$ and in the case of $\mathcal{M}_2$ to a new set

$$\mathcal{M}_2^k = \left\{ Q \in \mathcal{P}(\mathcal{Y}_k) \,\middle|\, \exists Q' \in \mathcal{M}_2 \text{ such that } \sum_{y_h} Q(y) = \sum_{y_h} Q'(y), \forall y \in \mathcal{Y} \right\}.$$

The set $\mathcal{M}_2^k$ captures the structure of the known variables after marginalization. If $\mathcal{M}_2$ is a graphical model then $\mathcal{M}_2^k$ corresponds to the conditional independence structure that results from marginalizing over the hidden states.

The algorithm iterates between the extended spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ on the left side of Figure 2.4. Using the following result, given in Theorem 2.5.1, we gain that this minimization is equivalent to an m-projection of the probability distribution of the observed data, $\tilde{P}$, to the marginalized set $\mathcal{M}_2^k$. Therefore, this algorithm results in the MLE, as shown in the equation (2.3).

If $\mathcal{M}_2$ is given by a chain graph model, as described in the Section 2.4, then $\mathcal{M}_2^k$ is a stratified exponential family corresponding to a chain mixed graph.

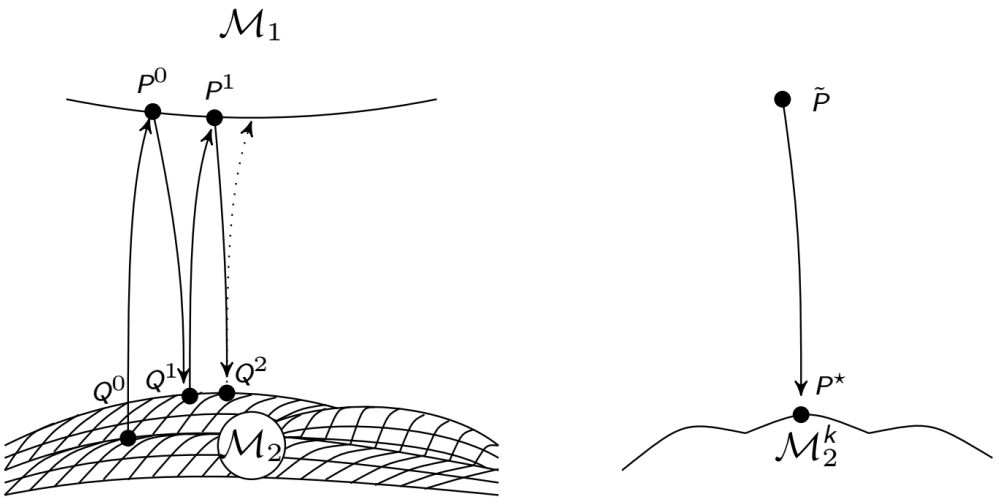


Figure 2.4: Sketch of Theorem 2.5.1.

Theorem 2.5.1 (Theorem 7 from Reference [Amari92]). *The minimum divergence between $\mathcal{M}_1$ and $\mathcal{M}_2$ is equal to the minimum divergence between $\tilde{P}$ and $\mathcal{M}_2^k$*

$$\inf_{P \in \mathcal{M}_1, Q \in \mathcal{M}_2} D_{\mathcal{Y}_k \times \mathcal{Y}_h}(P \parallel Q) = \inf_{Q_k \in \mathcal{M}_2^k} D_{\mathcal{Y}_k}(\tilde{P} \parallel Q_k).$$

Proof of Theorem 2.5.1. Let $P, Q \in \mathcal{P}(\mathcal{Y}_k \times \mathcal{Y}_h)$, using the chain-rule for KL-divergence, see Theorem 2.5.3 in [Cover06], leads to

$$D_{\mathcal{Y}_k \times \mathcal{Y}_h}(P \parallel Q) = D_{\mathcal{Y}_k}(P \parallel Q) + D_{\mathcal{Y}_h|\mathcal{Y}_k}(P \parallel Q),$$

with the conditional KL-divergence $D_{\mathcal{Y}_h|\mathcal{Y}_k}(P \parallel Q)$, see Definition 2. This results in

$$\begin{aligned}
\inf_{P \in \mathcal{M}_1, Q \in \mathcal{M}_2} D_{\mathcal{Y}_k \times \mathcal{Y}_h}(P \parallel Q) &= \inf_{P \in \mathcal{M}_1, Q \in \mathcal{M}_2} \left(D_{\mathcal{Y}_k}(P \parallel Q) + D_{\mathcal{Y}_h|\mathcal{Y}_k}(P \parallel Q) \right) \\
&= \inf_{P \in \mathcal{M}_1, Q \in \mathcal{M}_2} \left(D_{\mathcal{Y}_k}(\tilde{P} \parallel Q) + D_{\mathcal{Y}_h|\mathcal{Y}_k}(P \parallel Q) \right) \\
&= \inf_{Q \in \mathcal{M}_2} \left(D_{\mathcal{Y}_k}(\tilde{P} \parallel Q) + \inf_{P \in \mathcal{M}_1} D_{\mathcal{Y}_h|\mathcal{Y}_k}(P \parallel Q) \right) \\
&= \inf_{Q \in \mathcal{M}_2^k} D_{\mathcal{Y}_k}(\tilde{P} \parallel Q).
\end{aligned}$$

The equality between the first and second line above stems from the definition of $\mathcal{M}_1$. Since $P(y_k) = \tilde{P}(y_k)$ for every $P \in \mathcal{M}_1$ and the first addend considers the KL-divergence only w.r.t. $\mathcal{Y}_k$, this leads to the equality

$$D_{\mathcal{Y}_k}(P \parallel Q) = D_{\mathcal{Y}_k}(\tilde{P} \parallel Q), \quad \forall P \in \mathcal{M}_1.$$

Note that $\inf_{P \in \mathcal{M}_1} D_{\mathcal{Y}_h|\mathcal{Y}_k}(P \parallel Q) = 0$, since for every $Q \in \mathcal{M}_2$ the product of the distributions $\tilde{P}(Y_k)$ and $Q(Y_h|Y_k)$ lies in $\mathcal{M}_1$. □

2.5.2 Iterative Scaling

If $\mathcal{M}$ is an exponential family, then the MLE of an empirical distribution $\tilde{P}$ can be calculated by applying an iterative algorithm, called iterative scaling algorithm or Iterative Proportional Scaling (IPS). This is a widely known algorithm, first defined in the statistics literature by Deming and Stephan in 1940 in [Deming40] to estimate cell probabilities in contingency tables. It has been analyzed further for example in [Brown93; Csiszár75; Csiszár89] and is known in connection to optimal transport problems as Sinkhorn algorithm, see, for instance, [Berman20; Peyré19].

There exist various types of iterative scaling algorithms. In [Drton09a] and [Sullivant18] the authors describe a variant of the original algorithm that can be applied to every exponential family, proposed as generalized iterative scaling by Darroch and Ratcliff in [Darroch72]. There are further algorithms like the Sequential Conditional Generalized Iterative Scaling [Goodman02] or the improved iterative scaling, defined in [Pietra95].

Here we introduce the classical iterative scaling algorithm, which is defined for a specific kind of exponential family that includes graphical models and that we name "partition models" in Section 6.2. In this case the spanning functions of the vector space F in Definition 4 are defined in the following way. Let $\Pi(\mathcal{Y}) = \{\mathcal{Y}_1, \dots, \mathcal{Y}_k\}$, $k \in \mathbb{N}$ be a partition of $\mathcal{Y}$. Then the indicator functions for the elements of the partition

$$\delta_{\mathcal{Y}_1}(y) = \begin{cases} 1, & \text{if } y \in \mathcal{Y}_1 \\ 0, & \text{otherwise} \end{cases}$$

span the vector space of the exponential family $\mathcal{E}(\Pi)$. A partition model is an exponential family that is defined as above via a set of partitions $\Pi^{1,\dots,\nu} = \left\{\Pi^1(\mathcal{Y}), \dots, \Pi^\nu(\mathcal{Y})\right\}$, $\nu \in \mathbb{N}$, and it is denoted by $\mathcal{E}(\Pi^{1,\dots,\nu})$.

Now we can also define a set of distributions for which the marginals on the elements of a partition coincide with the marginals of the empirical distribution $\tilde{P}$

$$L^i = \left\{ P \in \mathcal{P}(\mathcal{Y}) \,\middle|\, P(\mathcal{Y}^k) = \tilde{P}(\mathcal{Y}^k),\, \forall \mathcal{Y}^k \in \Pi^i \right\},$$

with $P(\mathcal{Y}^k) := \sum_{y \in \mathcal{Y}^k} P(y)$.

Sets of this form are called linear families. In the case of a general exponential family $\mathcal{E}(F)$, the corresponding linear family is defined by the spanning vectors f_j of the vector space F and the condition

$$\sum_y P(y) f_j(y) = \sum_y \tilde{P}(y) f_j(y).$$

The iterative scaling algorithm works by iteratively projecting to different linear families L^i, as depicted in Figure 2.5. The e-projection of P to a linear family L^i is well-known, proven in for example Lemma 4.1 in [Csiszár04] and given by

$$P'(y) = P(y) \frac{\tilde{P}(\mathcal{Y}^k)}{P(\mathcal{Y}^k)},$$

for all $y \in \mathcal{Y}^k \in \Pi^i$.

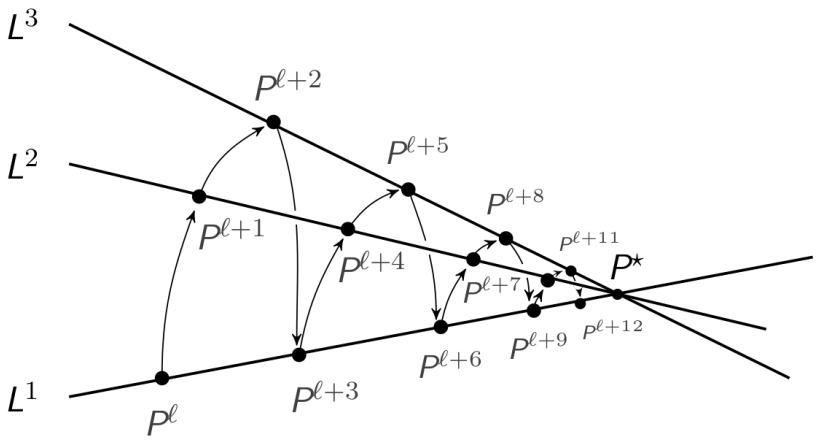


Figure 2.5: Sketch of the iterative scaling algorithm in the case of three linear families, published in [Coons24].

Now we are able to define the IPS for a partition model $\mathcal{E}(\Pi^{1,\dots,\nu})$. The starting point is the uniform distribution

$$P^0(y) = \frac{1}{|\mathcal{Y}|},$$

then a step of the algorithm is defined as

$$P^\ell(y) = P^{\ell-1}(y) \frac{\tilde{P}(\mathcal{Y}^k)}{P^{\ell-1}(\mathcal{Y}^k)},$$

for $y \in \mathcal{Y}^k \in \Pi^i$ and $i = \ell \bmod \nu$. A proof of the convergence can be found in Theorem 5.1 of [Csiszár04].

Let $\mathcal{L}$ be the non-empty intersection of the linear families, $\mathcal{L} = \bigcap_i L^i \neq \emptyset$. Then this algorithm minimizes the KL-divergence with respect to the first argument

$$P^\star = \arg\inf_{P \in \mathcal{L}} D(P \parallel Q),$$

for all $Q \in \overline{\mathcal{E}}(\Pi^{1,\dots,\nu})$.

This is equivalent to calculating the MLE with respect to $\overline{\mathcal{E}}(\Pi^{1,\dots,\nu})$, because of the equivalences in the following theorem.

Theorem 2.5.2 (Theorem 2.8 in [Ay17]). *For a probability distribution $P^\star \in \mathcal{P}(\mathcal{Y})$ the following statements are equivalent*

(1) $P^\star \in \mathcal{L} \cap \overline{\mathcal{E}}(\Pi^{1,\dots,\nu})$

(2) $P^\star \in \mathcal{L}$ *and* $D(P^\star \parallel Q) = \inf_{P \in \mathcal{L}} D(P \parallel Q)$ *for all* $Q \in \overline{\mathcal{E}}(\Pi^{1,\dots,\nu})$

(3) $P^\star \in \overline{\mathcal{E}}(\Pi^{1,\dots,\nu})$ *and* $D(P \parallel P^\star) = \inf_{Q \in \overline{\mathcal{E}}(\Pi^{1,\dots,\nu})} D(P \parallel Q)$ *for all* $P \in \mathcal{L}$.

There exists a unique distribution $P^\star$ that satisfies one and therefore all of these conditions.

The minimization with respect to the second argument, given in (3) of the above theorem, is equivalent to calculating the MLE, as shown in the equation (2.3). Therefore, this algorithm leads to the MLE.

Additionally, the minimization in number (2) in Theorem 2.5.2 gives us a different interpretation of the limit point of the iterative scaling algorithm. Since the chosen initial distribution is the uniform distribution, P^0, and it always lies in $\overline{\mathcal{E}}(\Pi^{1,\dots,\nu})$, we are able to rewrite the minimization in (2) to

$$\begin{aligned}\inf_{P \in \mathcal{L}} D(P \parallel P^0) &= \inf_{P \in \mathcal{L}} \sum_y P(y) \log \frac{P(y)}{\frac{1}{|\mathcal{Y}|}} \\ &= \log(|\mathcal{Y}|) + \sup_{P \in \mathcal{L}} \left(- \sum_y P(y) \log P(y) \right).\end{aligned}$$

Therefore this algorithm also leads to the element with the maximum entropy out of the elements in $\mathcal{L}$ and this is called the maximum entropy estimate. In [Jaynes57] the author formulates the maximum entropy principle, which states that in the case that we have a set of different probability distributions that fulfill certain desired restrictions, then we should choose the distribution that has the highest uncertainty, meaning the distribution with the highest entropy. This assures that the resulting distribution only assumes enough structure in order to satisfy the restrictions, but not more. Here the restrictions are given by the conditions for the linear families L^i.

This classical version of the iterative scaling algorithm, that we defined here, has the advantage that each iteration step produces a rational function. Hence this leads to a rational maximum likelihood estimator, which has been of recent interest [Coons21; Duarte21].

Additionally, we use this iterative scaling algorithm in the Chapters 4 and 5 in order to calculate measures for synergistic information and in Chapter 6.2 we define a sufficient condition for partition models that ensures that the iterative scaling algorithm converges in one cycle. This means that we only have to project to each linear family exactly once.

3 Complexity Measures in the Context of Integrated Information Theory

In this chapter we discuss different information geometric measures that have been proposed for calculating the Integrated Information in a stationary, time-homogeneous Markov process. We define two additional measures, the "ground truth Integrated Information", that includes a known exterior influence, and the "causal information integration" in which the exterior influence is unknown. The results in this chapter were largely published in [Langer20b] and the code of the calculated measures can be found at [Langer20a].

As described in the introduction the goal of this chapter is to find a suitable measure for Integrated Information to apply to embodied agents and thereby to relate the relevant information flow in the brain of the agent to its interaction with its environment. We do so in Chapter 4 and 5, where we use the ground truth Integrated Information to calculate the Integrated Information in the controller of simulated agents. The Markov processes discussed in this chapter will take the role of the controllers of the agents in the later chapters. Hence, to increase the consistency between the chapters we refer to the Markov process by $(C_t)_{t\in\mathbb{N}}$ for "controller".

First, we give an introduction to the Integrated Information Theory of consciousness in the next section.

3.1 The Integrated Information Theory of Consciousness

The main goal of the Integrated Information Theory (IIT) is to quantify the amount and quality of the consciousness of a system. Starting as a measure for brain complexity, published in [Tononi94] in 1994, this developed rapidly to a broad theory investigating the origin of consciousness. In this section we give a brief introduction to the history of the theory and its recent developments.

The above mentioned measure for brain complexity is defined in terms of the information theoretic concepts of mutual information and entropy. More details about this measure, its implications regarding the brain functions and suggestions for possible experimental tests can be found in [Tononi98b]. It is defined to have a high value if the cost of dividing a system into subsystems is high. Therefore, the initial idea behind Integrated Information is in agreement with our understanding of complexity, as addressed in Section 1.2.3. In particular, Tononi summarizes this idea in [Tononi08] on page 221 in the following way.

> "In short, integrated information captures the information generated by causal interactions in the whole, over and above the information generated by the parts."

In the first publications the concept of consciousness plays only a minor role, if mentioned at all. The explicit connection to consciousness appears in, for example, "Consciousness and Complexity" [Tononi98a] from 1998. There the authors characterize the type of mechanisms a group of neurons must perform in order to contribute to conscious experience. Two key properties are identified, the first one is a large amount of possible different states the physical system can be in, termed "Differentiation" here and "Information" in later publications. Secondly, the neurons should have a highly integrated structure, this property is named "Integration". More precisely, a system has a high functional integration if it cannot be divided into parts without loss of functionality. A group of neurons satisfying these condition is called a "dynamic core".

Instead of averaging over the different possible partitions of a system, as in [Tononi94], Tononi and Sporns propose in [Tononi03] to consider the minimum information bipartition,

which is the bipartition of a system with a minimal effective information. The effective information is given by the mutual information between the elements of the bipartition. A system that is not a subset of a system with a higher Integrated Information value is called a complex. Furthermore, in this publication the authors coin the symbol Φ for Integrated Information because the "I" symbolizes the information and the "⬭" represents the system the information is integrated in.

These concepts are further elaborated in [Tononi04] with the help of a thought experiment in which Tononi considers a high resolution camera taking a picture opposed to a human looking at the same scene. The camera can differentiate between the many states the photodiodes can be in but the different parts of the system do not interact with each other. Thereby he highlights that the information gets integrated inside the brain of a conscious observer and stays separate inside the camera.

An additional aspect of this theory is that it aims at not only quantifying the amount but also the quality of consciousness. To that end the conscious experience is equated with a structure, called quale, in the qualia space, which is an abstract relational space [Tononi04; Tononi08].

A summary of his theory can be found in "Consciousness as Integrated Information: a Provisional Manifesto" [Tononi08]. Here, the measure for calculating Integrated Information is still based on information theoretic quantities, such as entropy and mutual information. The difference between a full and a split system is calculated using the KL-divergence similar to the approach discussed in this thesis.

The authors of [Balduzzi08] extended this theory from stationary stochastic processes to discrete dynamical systems. Their measure for integrated information is discussed and generalized in [Barrett11], which results in a measure similar to the stochastic interaction, defined in [Ay15a], that we introduce and discuss in Section 3.2.1. Comparisons of these types of measures are given in, for example, [Oizumi16b; Kanwal17; Mediano18].

A central aspect of the next iteration of the theory, called IIT 3.0, is to start with a phenomenological perspective on consciousness by considering thought experiments, which lead to axioms, as discussed in [Tononi12; Oizumi14]. This approach leads to five axioms, as defined in [Oizumi14]:

- "EXISTENCE: Consciousness exists – it is an undeniable aspect of reality."
- "COMPOSITION: Consciousness is compositional (structured): each experience consists of multiple aspects in various combinations."
- "INFORMATION: Consciousness is informative: each experience differs in its particular way from other possible experiences."
- "INTEGRATION: Consciousness is integrated: each experience is (strongly) irreducible to non-interdependent components."
- "EXCLUSION: Consciousness is exclusive: each experience excludes all others – at any given time there is only one experience having its full content, rather than a superposition of multiple partial experiences; each experience has definite borders – certain things can be experienced and others cannot;"

From these axioms the authors derive mechanisms or postulates, which they then translate to mathematical expressions. Therefore, the calculus for the Integrated Information

value is fundamentally different and significantly more complex compared to previous versions. The core idea, calculating the difference between a full system and a partial one, is still valid but the definition of a partial system consists of multiple steps. There they for example differentiate between causes and effects in the system. Additionally, they use the Wasserstein metric, also known as earth movers distance, instead of the KL-divergence. This metric not only takes into account the probability distributions but also the distance between the states. The complexity of this new formulation does not allow it to be applied to large systems. Hence, the authors of [Marshall16] compare the results of this formulation of IIT with a measure for state differentiation. There they explore the possibility to use the state differentiation as a proxy for Integrated Information.

This theory can be applied to a system at various spatial and temporal scales. In [Hoel16] the authors discuss simplified neuronal systems and show that the macro level can have a higher Integrated Information compared to the smaller scales. The discussion is continued in [Marshall18].

Throughout the development of this theory the proponents always reference the properties of the structure of the human brain. Examples for studies of Integrated Information using data from EEG scans can be found in [Casali13; Kim18; Massimini10; Massimini05]. Overviews over a more neurological and neurobiological perspective on consciousness, including IIT, are given in [Koch16; Tononi16a]. In the opinion article [Tononi16b] Tononi highlights different aspects of IIT in relation to structures in the brain.

In [Kleiner21] Kleiner and Tull introduce a mathematically rigorous, axiomatic approach that summarizes IIT 3.0 as well as other related branches, such as the recently developed quantum IIT [Zanardi18; Albantakis23]. In quantum IIT the proponents aim to extend the IIT framework to quantum systems. These publications already refer to the latest version of the theory, IIT 4.0.

The phenomenon of spatial experience from an IIT perspective is the subject of the analysis in the publication [Haun19]. There, the authors refer to IIT 3.0 but introduce "existence" as a starting point rather than an axiom. Instead the first axiom is now called "intrinsicality". This describes the observation that conscious experience is inherently subjective. This intrinsicality plays an important role in [Albantakis19]. There the authors show that systems with the same global dynamics may differ in their internal composition and therefore have an entirely different attached phenomenology from the IIT perspective. The calculus of the Integrated Information measure is changed here again and returns to using the KL-divergence.

This is formalized in terms of a measure for intrinsic information in [Barbosa20], which is derived from three properties, namely "causality", "intrinsicality" and "specificity". In [Barbosa21] the authors argue that this measure captures all the axioms of IIT and therefore is termed the unique measure for IIT in the preprints [Marshall22] and [Albantakis22]. The latter is called "Integrated Information Theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms".

Claiming that any version of this theory determines consciousness exactly is highly controversial. Aaronson famously criticizes in his blog post [Aaronson14b] the definition of the Integrated Information measure and uses expander graphs to define an example with a high Integrated Information value that he would not consider to be conscious. The surprising answer by Tononi, discussed in [Aaronson14a], is to claim that these examples indeed have a consciousness. He goes even further and explains that a simple grid of XOR gates can have a high Integrated Information, therefore have a consciousness but a qualitative different one compared to humans. This leads IIT in the direction of panpsychism, which claims that everything in the universe is conscious.

The measure is also criticized in [Merker22] where the authors claim that it simply

measures the global information transfer in differentiated networks. They call this the "network efficiency".

Furthermore, there exists some debate regarding the validity of the axioms. In the opinion paper [Bayne18] the author discusses all the axioms of IIT 3.0 and concludes that the axioms fail their purpose. Cerullo addresses mainly the fifth axiom "exclusion" and discards it as not self-evident, [Cerullo15]. Pautz asks in [Pautz19] the more fundamental question of what exactly a certain amount of consciousness is.

The authors of [Mediano22b] highlight different strengths and problems of the theory and advocate for a weaker version that focuses on explaining features of consciousness, not the phenomenon itself. Another overview of different aspects of IIT can be found in [Mallatt21] where IIT is compared to neurobiological naturalism. This describes the theory in which consciousness is an evolved emergent feature of natural systems.

In this thesis we follow the suggestion from the authors of [Mediano22a] and adopt a more pragmatic point of view. Here, we do not consider Integrated Information to be a measure of consciousness but to quantify the complexity of a system. This system is given by the controller of an artificial agent.

The branch of Integrated Information measures we are interested in is defined in connection with information geometry. We follow the strategy of Oizumi et al. [Oizumi16b] and Amari et al. [Amari18] and restrict attention to measuring the Integrated Information in discrete n-dimensional stationary Markov processes from an information geometric point of view. These measures are introduced in the next section. Additionally, we define two measures, published in [Langer20b], that quantify the Integrated Information depending on whether the system is subjected to known or unknown external influences or not. The former one is then applied to artificial agents in Chapter 4 and 5.

3.2 Integrated Information Measures

The measures corresponding to Integrated Information discussed in this chapter analyze the information flow in a system from a time t to $t+1$.

The systems are modeled as discrete, stationary, n-dimensional Markov processes $(C_t)_{t\in\mathbb{N}}$

$$C_t = (C_t^1, \dots, C_t^n), \qquad C_{t+1} = (C_{t+1}^1, \dots, C_{t+1}^n)$$

and we define $B = (C_t, C_{t+1})$ on a finite set $\mathcal{B} \neq \emptyset$, which is the Cartesian product of the state spaces of C_t^j, $j \in \{1, \dots, n\}$, denoted by $\mathcal{C}^j$

$$\mathcal{B} = \mathcal{C} \times \mathcal{C} = \bigtimes_{i=j}^{n} \mathcal{C}^j \times \bigtimes_{j=1}^{n} \mathcal{C}^j.$$

The information flow from t to $t+1$ is represented by the connections from the nodes C_t^j to the nodes $C_{t+1}^{j'}$ in $j, j' \in \{1, \dots, n\}$, as displayed in Figure 3.1.

Since we assume that the process is Markovian, stationary and time-homogeneous, we are able to restrict the following discussion to one timestep.

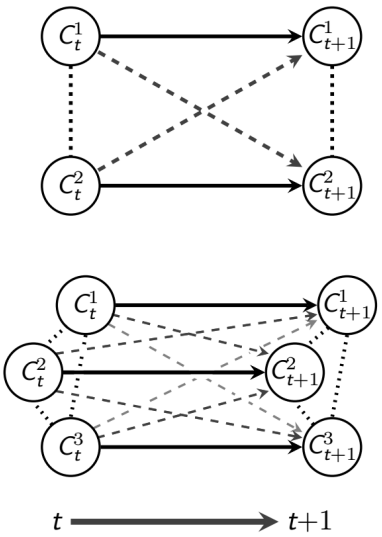


Figure 3.1: The fully connected system for $n = 2$ and $n = 3$.

We denote the set of probability distributions belonging to the Markov processes defined above by $MP(\mathcal{C})$.

Denote the complement of C_t^j in C_t by

$$C_t^{J\setminus\{j\}} = (C_t^1, \ldots, C_t^{j-1}, C_t^{j+1}, \ldots, C_t^n) \text{ with } J = \{1, \ldots, n\}.$$

Corresponding to this notation $c_t^{J\setminus\{j\}} \in \mathcal{C}_t^{J\setminus\{j\}}$ describes the elementary events of $C_t^{J\setminus\{j\}}$ and to simplify we write $b \in \mathcal{B}$ instead of $(c_t, c_{t+1}) \in \mathcal{C} \times \mathcal{C}$.

The Integrated Information measures discussed in this thesis share one core idea: they measure the difference between the initial system and one with no information integration. We call the former a "full" system, because it has all possible connections between the nodes, and the latter a "split" system. The graphical representations of full systems in the case of $n = 2, 3$ are depicted in Figure 3.1. Graphical models are a tool to describe the conditional independence structure of sets of probability distributions. As described in Section 2.4 in more detail, graphs and conditional independence statements provide the foundation of Pearls causality theory, [Pearl85a; Pearl09]. Considering this, we argue that the split system that is used to define a measure of causal connections should have a graphical representation.

Given a specific split system the difference between the measures corresponding to the full and split systems is calculated by using the KL-divergence, as suggested in [Ay01; Ay15a]. This is described in the following definition.

Definition 14 (Complexity)**.** Let $\hat{P}$ be the distribution of the fully connected system and $\mathcal{M}$ be the set of probability distributions on $\mathcal{B}$ that corresponds to a split system. Then we minimize the KL-divergence between $\mathcal{M}$ and $\hat{P}$ to calculate the complexity

$$\Phi_{\mathcal{M}} = \inf_{Q\in\mathcal{M}} D_{\mathcal{B}}(\hat{P} \parallel Q) = \sum_{b\in\mathcal{B}} \hat{P}(b) \log \frac{\hat{P}(b)}{Q(b)}.$$

Since minimizing the KL-divergence with respect to the second argument is called m-projection, as defined in Section 2.5.1, we call $P^\star$ with

$$P^\star = \arg\inf_{Q\in\mathcal{M}} D_{\mathcal{B}}(\hat{P} \parallel Q)$$

the projection of $\hat{P}$ to $\mathcal{M}$.

The different measures for Integrated Information discussed in this section result from the various ways in which we can define a split system. The goal is to measure the information that gets integrated between different nodes in different points in time. These connections are the dashed connections in Figure 3.1, those are also called cross-influences in Reference [Oizumi16b]. In order to emphasize the causal aspect of our approach we refer to them as causal cross-connections.

The authors of [Oizumi16b] and [Amari18] postulate two properties that a valid measure for Integrated Information should satisfy. They ensure that the causal cross-connections are removed in the split system by requiring C_{t+1}^j to be independent of $C_t^{j'}$ given $C_t^{J\setminus\{j'\}}$, $j' \neq j$. This is expressed in the following property.

Postulate 1. *A split system that is used to define an Integrated Information measure should satisfy the Markov condition*

$$Q(C_t^{j'}, C_{t+1}^j \mid C_t^{J\setminus\{j'\}}) = Q(C_t^{j'} \mid C_t^{J\setminus\{j'\}}) Q(C_{t+1}^j \mid C_t^{J\setminus\{j'\}}), \quad j' \neq j, \tag{3.1}$$

with $Q \in \mathcal{P}(\mathcal{B})$. This can also be written as

$$C_{t+1}^j \perp\!\!\!\perp C_t^{j'} | C_t^{J\setminus\{j'\}}.$$

Now we focus on the remaining connections. The dotted lines are connections between nodes in the same point in time. Since we are interested in the information flow between t and $t+1$, the distribution in t, including the connection between the C_t^js, should stay unchanged in the split system.

There are two different mechanisms that lead to the same-time connections between the C_{t+1}^js. Firstly, they might result from common internal influences, more precisely, a correlation between the C_t^js can be passed on to the next point in time. Secondly, in Section 6.9 in [Amari16] Amari points out that there could be a common exterior influence on the C_{t+1}^js. Although the Integrated Information measure is defined to assess the internal information flow independently of external influences the whole system is in general not completely independent of its environment.

These dotted connections between the C_{t+1}^js play an important role in the definition of the second postulate from [Oizumi16b; Amari18]. There they discuss the split system in which all the solid and dashed connections are removed. The solid arrows represent the influence of a node in t on itself in $t+1$. Removing the solid and dashed connections at the same time results in a system in which the different points in time are completely disconnected as shown in Figure 3.2. The distributions factoring according to these split systems are

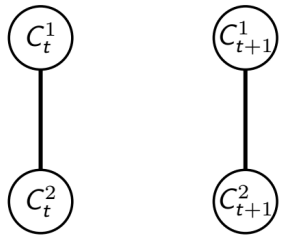


Figure 3.2: The graph corresponding to $\mathcal{M}_I$.

$$\mathcal{M}_I = \{Q \in \mathcal{P}^\circ(\mathcal{B}) | \, Q(b) = Q(c_t)Q(c_{t+1}), \forall b = (c_t, c_{t+1}) \in \mathcal{B}\}$$

and the measure Φ_I is given by the mutual information $I(C_t; C_{t+1})$, which is defined in Definition 3.

This measure calculates the difference between the full system and one in which there is no information flow between the time steps. Hence the authors of [Oizumi16a; Oizumi16b; Amari18] argue that Φ_I should be an upper bound for all Integrated Information measures.

Postulate 2. *The mutual information should be an upper bound for an Integrated Information measure*

$$\Phi_{\mathcal{M}} = \inf_{Q \in \mathcal{M}} D_{\mathcal{B}}(\hat{P} \mid Q) \leq I(C_t; C_{t+1}) = \Phi_I.$$

This postulate led to a discussion in [Kanwal17] about the treatment of the connection between the C_{t+1}^js. As discussed earlier, there are two different mechanisms leading to this edge. On the one hand this connection takes into account that there might exist a common exterior influence affecting all the C_{t+1}^js, as pointed out in [Amari16]. This is depicted on the top in Figure 3.3 by the additional node W_C. Considering that we only want to measure the dashed connections this should not be counted as part of the Integrated Information.

On the other hand the correlations between the C_t^js can be passed on to the C_{t+1}^js via the solid and dashed arrows. When we calculate the mutual information we create the edge between the C_{t+1}^js by marginalizing to C_{t+1}. The distribution that results from this also contains these passed on correlation, as depicted on the bottom of Figure 3.3. Kanwal et al. discuss the question of how much of these correlations integrate information in the system and should therefore be measured in [Kanwal17]. There they differentiate between intrinsic and extrinsic influences that cause the edges between the C_{t+1}^js. This edge between the C_{t+1}^js includes the influences of the solid arrows and common exterior influences as intended but might also compensate for part of the dashed causal cross-connections.

In [Kanwal17] the authors discuss an example of a system without a common exterior influence. There the edge between the C_{t+1}^js only consists of influences from the solid and

dashed connections. They show that there are cases in which a measure that is sure to only remove the causal cross-connections can exceed Φ_I. This means that there exists a split distribution with the edge between the C^j_{t+1}s that is closer to $\hat{P}$ compared to the distributions that only lack the causal cross-connections. Hence, the undirected edge between the C^j_{t+1}s compensates for a part of the causal cross-connections. This leads to the conclusion that Φ_I does not always measure all the intrinsic causal cross-connections. Therefore Kanwal et al. question the use of the mutual information as an upper bound in Postulate 2.

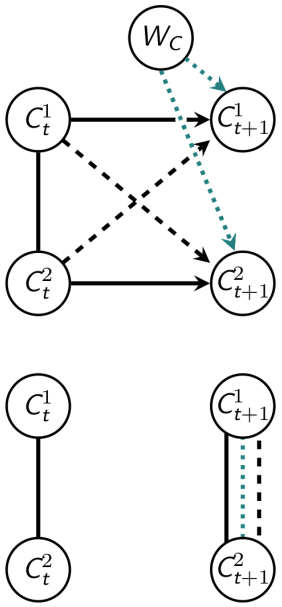


Figure 3.3: Influences on C_{t+1} in the full and the split system corresponding to Φ_I.

We would like to contribute a different perspective to this discussion and distinguish between the situations with and without unknown exterior influences. If we do not have any unknown influences on our system, then using the mutual information as an upper bound is unreasonable, as Kanwal et al. showed.

However, there are good arguments for using Postulate 2 in the case of unknown influences. Requiring the condition in Postulate 2 to hold does not necessarily lead to split model in which there is a fixed connection between the C^j_{t+1}s. It can instead be seen as forcing $\mathcal{M}_I$ to be a subset of the set of split distributions. This is a slightly different perspective and the measures Φ_{CIS} and Φ_{CII} satisfy Postulate 2 in this way. Although Φ_I does not measure all the intrinsic influences in general the following argument supports still using Postulate 2.

Consider a full system with the distribution $\hat{P}(b) = \hat{P}(c_t)\hat{P}(c_{t+1})$, $\forall b \in \mathcal{B}$. This system has a common exterior influence on the C^j_{t+1}s and no information flow between the different points in time. Therefore a measure for Integrated Information $\Phi_{\mathcal{M}}$ should be zero for all distributions of this form. This is the case if and only if $\mathcal{M}_I \subseteq \mathcal{M}$. In order to emphasize this point we propose a modified version of Postulate 2.

Postulate 3. *For every Integrated Information measure $\Phi_{\mathcal{M}}$ the set $\mathcal{M}_I$ should be a subset of the split model $\mathcal{M}$.*

Note that if Postulate 3 holds, then

$$\Phi_{\mathcal{M}} = \inf_{Q \in \mathcal{M}} D_{\mathcal{B}}(\hat{P} \mid Q) \leq I(C_t; C_{t+1}).$$

Hence Postulate 2 is a consequence of Postulate 3. Every Integrated Information measure discussed in this thesis that satisfies Postulate 2 also fulfills Postulate 3. To simplify we keep referring to Postulate 2 for the remainder of this thesis. Now we are able to discuss three different Integrated Information measures, namely stochastic interaction, geometric Integrated Information and CIS Integrated Information.

3.2.1 Stochastic Interaction

The complexity measure called "stochastic interaction" was introduced by Ay in [Ay01] in 2001, later published in [Ay15a] and Barrett and Seth discuss it in [Barrett11] in the context of Integrated Information. The core idea of this measure is to quantify how much the whole is more then the sum of its parts. However, it does not consider any exterior influence and therefore does not fulfill the property in Postulate 2.

In this case the split system only allows the connections among the random variables in t and additionally the connections between C_t^j and C_{t+1}^j. This leads to a split system where only the same random variable in different points in time are connected, corresponding to the solid arrows in 3.1, and the variables in time t. A graphical representation for $n = 2$ can be found in Figure 3.4.

Definition 15 (Stochastic Interaction)**.** The set of distributions belonging to the split model in the sense of Stochastic Interaction can be defined as

$$\mathcal{M}_{SI} = \left\{ Q \in \mathcal{P}^\circ(\mathcal{B}) \,\middle|\, Q(C_{t+1} \mid C_t) = \bigotimes_{j=1}^{n} Q(C_{t+1}^j \mid C_t^j) \right\}$$

and the complexity measure can be calculated as follows

$$\Phi_{SI} = \inf_{Q \in M_{SI}} D_{\mathcal{B}}(\hat{P} \parallel Q) = \sum_{j=1}^{n} H_{\hat{P}}(C_{t+1}^j \mid C_t^j) - H_{\hat{P}}(C_{t+1} \mid C_t),$$

as shown [Ay15a].

This does not satisfy Postulate 2, as Amari points out in [Amari16]. Hence, this measure is not applicable in the case of an exterior influences on the C_{t+1}^js. Such an influence can cause the C_{t+1}^js to be correlated even in the case of independent C_t^js and no causal cross-connections.

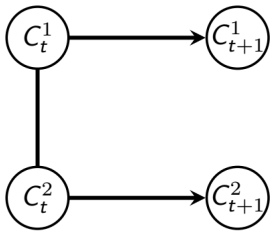


Figure 3.4: The graph corresponding to $\mathcal{M}_{SI}$.

Consider a setting without exterior influences then Φ_{SI} quantifies only the strength of the causal cross-connections and is therefore a reasonable choice for an Integrated Information measure. In this situation we do not require the measure to satisfy Postulate 2. Accounting for an exterior influence on the C_{t+1}^js that does not exist leads to a split system which compensates for a part of the removed causal cross-connections. Then the resulting measure does not quantify all of the interior causal cross-connections, as discussed earlier in the context of Postulate 2.

To force the model to satisfy Postulate 2 one can add the interaction between the nodes C_{t+1}^j and $C_{t+1}^{j'}$, which results in the measure geometric Integrated Information [Amari16].

3.2.2 Geometric Integrated Information

In [Amari16] Amari introduces an additional edge to the split system of stochastic interaction. This edge is between the C_{t+1}^js and guarantees that this new measure, called geometric Integrated Information, does satisfy Postulate 2.

Definition 16 (Geometric Integrated Information)**.** The graphical model corresponding to the graph in Figure 3.5 is the set

$$\mathcal{M}_G = \left\{ P \in \mathcal{P}^\circ(\mathcal{B}) \,\middle|\, \exists f_1, \ldots, f_{n+2} \in \mathbb{R}_+^{\mathcal{B}} \text{ s.t. } P(b) = f_{n+1}(c_t) f_{n+2}(c_{t+1}) \prod_{j=1}^{n} f_j(c_t^j, c_{t+1}^j) \right\}$$

and the measure is defined as

$$\Phi_G = \inf_{Q \in \mathcal{M}_G} D_{\mathcal{B}}(\hat{P} \parallel Q).$$

The set $\mathcal{M}_G$ is also called the diagonally split model in [Amari18]. The manifold $\mathcal{M}_I$ is a subset of $\mathcal{M}_G$, hence

$$\Phi_G \leq \Phi_I \tag{3.2}$$

and the measures satisfies Postulate 2.

Note that at first glance $\mathcal{M}_G$ defines a set of probability distributions that factor according to the undirected graph on the bottom of Figure 3.5. This is the case because both graphs are Markov equivalent as one can easily see using the criterion given by Frydenberg in [Frydenberg90] and discussed in Section 2.4.3. The chain components are $\{C_t^1, \dots, C_t^n\}$ and $\{C_{t+1}^1, \dots, C_{t+1}^n\}$. The boundary of the first component is empty and the boundary of the second component consists of the first one and is therefore complete, meaning that both graphs are Markov equivalent. Hence the factorizations corresponding to these graphs are also equivalent. For more details see Section 2.4.3.

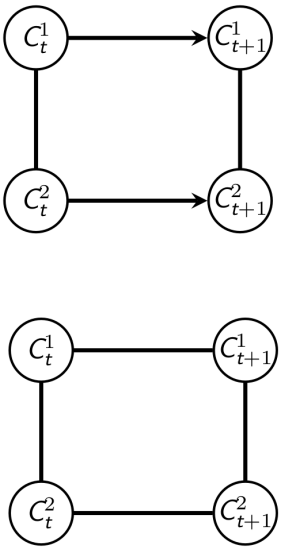


Figure 3.5: The graph corresponding to $\mathcal{M}_G$ on the top and the Markov equivalent undirected graph on the bottom.

Introducing this additional edge has a caveat. The new model is not causally split in the sense that the corresponding distributions in general do not satisfy Postulate 1. This can be verified by analyzing the conditional independence structure of the graph as described in Section 2.4. The moral graph of $An(C_t^1, C_t^2, C_{t+1}^1)$ is the fully connected graph because every C_t^j is connected to the chain component in which every C_{t+1}^j lies. Therefore the global chain graph Markov Property in Definition 9 is not satisfied.

An additional problem is that the fixed edges between the C_{t+1}^js might lead to these connections being stronger than they originally are in the distribution of the fully connected system $\hat{P}$. A result is that in some cases an effect of the causal cross-connections gets atoned for by the new edge. We discussed this in the context of Postulate 2.

This measure has no closed form solution but because this graph is Markov equivalent to an undirected graph $\mathcal{M}_G$ is an exponential family. Hence, we are able to calculate the corresponding split system with the help of the iterative scaling algorithm. There we iteratively project to linear families that are given by fixing the values of the edges

$$\hat{P}(C_t),\ \hat{P}(C_{t+1}), \hat{P}(C_t^j, C_{t+1}^j), \quad j \in \{1, \dots, n\}$$

using e-projections. This is described in more detail in 2.5.2.

3.2.3 CIS Integrated Information

The first measure that satisfies both postulates defines the split model not via a graphical model but by using conditional independence statements, so we denote it by Φ_{CIS} and call it CIS Integrated Information. It is called "Integrated Information" in [Oizumi16b] and its model is referred to as "Causally split model" in Reference [Amari18]. This measure is derived from the first postulate by requiring C_{t+1}^j to be independent of $C_t^{J\setminus\{j\}}$ given C_t^j. More details to CIS can be found in Section 2.4.

Definition 17 (CIS Integrated Information)**.** The set of distributions that belong to the split system corresponding to CIS Integrated Information is defined as

$$\mathcal{M}_{CIS} = \left\{ Q \in \mathcal{P}^\circ(\mathcal{B}) \mid Q(C_{t+1}^j \mid C_t) = Q(C_{t+1}^j \mid C_t^j), \text{ for all } j \in \{1, \dots, n\} \right\} \tag{3.3}$$

and this leads to the measure

$$\Phi_{CIS} = \inf_{Q \in \mathcal{M}_{CIS}} D_{\mathcal{B}}(\hat{P} \parallel Q).$$

The requirements for the distributions in (3.3) can be written as conditional independent statements:

$$C_{t+1}^j \perp\!\!\!\perp C_t^{J\setminus\{j\}} \mid C_t^j.$$

Unfortunately, these conditional independence statements do not have a graphical representation in general. This measure is strongly connected to Postulate 1 in the following way.

Proposition 2. *Applying the conditional independence statements in (3.1)*

$$Q(C_t^j, C_{t+1}^{j'} \mid C_t^{J\setminus\{j\}}) = Q(C_t^j \mid C_t^{J\setminus\{j\}}) Q(C_{t+1}^{j'} \mid C_t^{J\setminus\{j\}}), \quad j \neq j'$$

to all pairs $j, j' \in \{1, \dots, n\}$ leads to a probability distribution for which

$$Q(C_{t+1}^j | C_t) = Q(C_{t+1}^j | C_t^j) \tag{3.4}$$

holds for all $j \in J$. Hence, this distribution lies in $\mathcal{M}_{CIS}$.

Proof. For $n = 2$ this is immediate. Let now $n \geq 3$ and $i, j, k \in J = \{1, \dots, n\}$ with $i \neq j \neq k \neq i$. Applying (3.1) two times leads to

$$\begin{aligned}
Q(c_{t+1}^j, c_t) &= \frac{Q(c_{t+1}^j, c_t^{J\setminus\{i\}}) Q(c_t)}{Q(c_t^{J\setminus\{i\}})} \\
Q(c_{t+1}^j, c_t) &= \frac{Q(c_{t+1}^j, c_t^{J\setminus\{k\}}) Q(c_t)}{Q(c_t^{J\setminus\{k\}})} \\
Q(c_{t+1}^j, c_t^{J\setminus\{i\}}) Q(c_t^{J\setminus\{k\}}) &= Q(c_{t+1}^j, c_t^{J\setminus\{k\}}) Q(c_t^{J\setminus\{i\}})
\end{aligned}$$

for all $(c_t, c_{t+1}^j) \in \mathcal{C} \times \mathcal{C}^j$. Marginalizing over the elements of $\mathcal{C}^k$ yields

$$\begin{aligned}
Q(c_{t+1}^j, c_t^{J\setminus\{i,k\}}) Q(c_t^{J\setminus\{k\}}) &= Q(c_{t+1}^j, c_t^{J\setminus\{k\}}) Q(c_t^{J\setminus\{i,k\}}) \\
Q(c_{t+1}^j | c_t^{J\setminus\{i,k\}}) &= Q(c_{t+1}^j | c_t^{J\setminus\{k\}}).
\end{aligned}$$

Using inductively the remaining relations results in (3.4). □

Hence, every model satisfying Postulate 1 is a subset of $\mathcal{M}_{CIS}$. Since $\mathcal{M}_I$ fulfills the CISs of Postulate 1, the relation $\mathcal{M}_I \subseteq \mathcal{M}_{CIS}$ holds and Φ_{CIS} satisfies Postulate 2.

In order to show that Φ_{CIS} also satisfies Postulate 1 we rewrite the condition in Postulate 1 as

$$Q(C_{t+1}^j | C_t) = Q(C_{t+1}^j | C_t^{J\setminus\{j'\}}), \text{ for } j, j' \in J \text{ and } j' \neq j.$$

The definition of $\mathcal{M}_{CIS}$ leads to

$$Q(C_{t+1}^j | C_t) = Q(C_{t+1}^j | C_t^j) = Q(C_{t+1}^j | C_t^{J\setminus\{j'\}}),$$

for $Q \in \mathcal{M}_{CIS}$, $j, j' \in J$ and $j' \neq j$. Therefore Φ_{CIS} satisfies Postulate 1.

This measure does not have a closed form solution in general. In [Oizumi16b] the authors derive an analytical solution for Gaussian variables and use Newton's method in the case of discrete variables.

Another caveat of this measure is the lack of a graphical representation. It not only makes the calculation of this measure more difficult, but it also complicates interpreting the causal nature of the elements of $\mathcal{M}_{CIS}$. Since we only know that they satisfy the conditional independence statements in Postulate 1, this does not lead to a parametrization of the distributions. Regarding the distribution $P^\star$ in $\mathcal{M}_{CIS}$, that minimizes the following KL-divergence, we are able to write

$$\begin{aligned} P^\star &= \operatorname*{arg\,inf}_{Q \in \mathcal{M}_{CIS}} D_{\mathcal{B}}(\hat{P} \parallel Q) \\ &= \operatorname*{arg\,inf}_{Q \in \mathcal{M}_{CIS}} \Big(D_{\mathcal{C}}(\hat{P}(C_t) \parallel Q(C_t)) + D_{\mathcal{C}|\mathcal{C}}(\hat{P}(C_{t+1}|C_t) \parallel Q(C_{t+1}|C_t)) \Big). \end{aligned}$$

Now, since the conditions in $\mathcal{M}_{CIS}$ do not restrict $Q(C_t)$ in any way, but only apply to $Q(C_{t+1}|C_t)$, we are always able to choose an element $P^\star$ from $\mathcal{M}_{CIS}$ for which the equality $P^\star(C_t) = \hat{P}(C_t)$ holds. This leads to

$$P^\star = \operatorname*{arg\,inf}_{Q \in \mathcal{M}_{CIS}} D_{\mathcal{C}|\mathcal{C}}(\hat{P}(C_{t+1}|C_t) \parallel Q(C_{t+1}|C_t)).$$

Therefore we know that the marginals of $P^\star$ on C_t equal the marginals of the distribution of the fully connected system, $\hat{P}(C_t)$, but we have no further information on the structure of $P^\star(C_{t+1}|C_t)$.

3.3 Ground Truth Integrated Information

In this section we propose a measure for Integrated Information that we define using the ground truth.

As discussed in the context of Postulate 2, there might exist exterior influences on the controller nodes. In Chapter 4 we see an example of a situation in which all the influences on the controller of the agent are known.

In this case there are no unknown exterior influence and therefore we can represent the combined exterior influences on the controller nodes by a variable W_C. This is shown in Figure 3.6 on the top.

If we have the distribution of the whole model we are able to extend the concepts discussed in the previous section to the larger space $\mathcal{B} \times \mathcal{W}_C$, where $\mathcal{W}_C$ is the state space of W_C. This allows us to account for the influences from the environment directly and thereby to define an Integrated Information measure in which we really only remove the causal cross-connections as shown in Figure 3.6 on the bottom. Thus we can interpret this measure as the true measure of Integrated Information if all the influences on the systems are available and therefore we call it ground truth Integrated Information, Φ_T.

Note that using the measure Φ_{SI} in the setting with no external influences can be seen as a special case of Φ_T in which the controller nodes are independent of the exterior influences in W_C.

The set of distributions belonging to the larger, fully connected model is called $\mathcal{M}_T^f$, which is an exponential family on the full space. On the bottom of Figure 3.6 is the graph corresponding to the split system, which is denoted by $\mathcal{M}_T$.

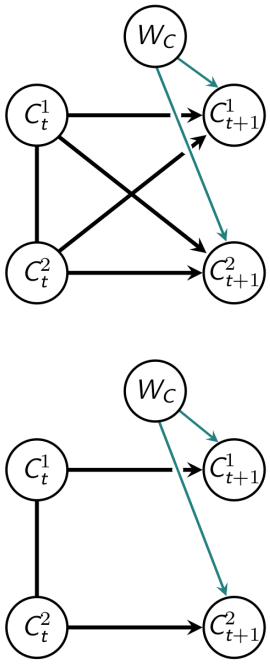


Figure 3.6: The graphs corresponding to $\mathcal{M}_T^f$ (top) and $\mathcal{M}_T$ (bottom).

$$\mathcal{M}_T^f = \left\{ P \in \mathcal{P}^\circ(\mathcal{B}\times\mathcal{W}_C) \,\middle|\, P(b,w) = P(c_t)\prod_{j=1}^{n} P(c_{t+1}^j|c_t,w)P(w),\ \forall (b,w)\in\mathcal{B}\times\mathcal{W}_C \right\}$$

$$\mathcal{M}_T = \left\{ P \in \mathcal{P}^\circ(\mathcal{B}\times\mathcal{W}_C) \,\middle|\, P(b,w) = P(c_t)\prod_{j=1}^{n} P(c_{t+1}^j|c_t^j,w)P(w),\ \forall (b,w)\in\mathcal{B}\times\mathcal{W}_C \right\}.$$

Calculating the m-projection of $P \in \mathcal{M}_T^f$ to $\mathcal{M}_T$ results in the new measure.

Proposition 3. *Let $\hat{P} \in \mathcal{M}_T^f$. Minimizing the KL-divergence between $\hat{P}$ and $\mathcal{M}_T$ leads to*

$$\Phi_T = \inf_{Q\in\mathcal{M}_T} D_{\mathcal{B}\times\mathcal{W}_C}(\hat{P}\parallel Q) = \sum_{b,w}\hat{P}(b,w)\,\log\frac{\prod_{j=1}^{n}\hat{P}(c_{t+1}^j|c_t,w)}{\prod_{j=1}^{n}\hat{P}(c_{t+1}^j|c_t^j,w)}$$
$$= \sum_{j=1}^{n} I(C_{t+1}^j; C_t^{J\setminus\{j\}}|C_t^j, W_C).$$

Proof of Proposition 3. Let $P \in \mathcal{M}_T^f$ and $Q \in \mathcal{M}_T$. Then the KL-divergence between P and Q can be bound from below as follows

$$D_{\mathcal{B}\times\mathcal{W}_C}(P\parallel Q) = \sum_{b,w} P(b,w)\,\log\frac{P(c_t)\prod_{j=1}^{n}P(c_{t+1}^j|c_t,w)P(w)}{Q(c_t)\prod_{j=1}^{n}Q(c_{t+1}^j|c_t^j,w)Q(w)}$$

$$= \sum_{c_t} P(c_t)\,\log\frac{P(c_t)}{Q(c_t)} + \sum_{b,w}P(b,w)\,\log\frac{\prod_{j=1}^{n}P(c_{t+1}^j|c_t,w)}{\prod_{j=1}^{n}Q(c_{t+1}^j|c_t^j,w)} + \sum_{w}P(w)\,\log\frac{P(w)}{Q(w)}$$

$$\geq \sum_{c_t} P(c_t)\,\log\frac{P(c_t)}{P(c_t)} + \sum_{b,w}P(b,w)\,\log\frac{\prod_{j=1}^{n}P(c_{t+1}^j|c_t,w)}{\prod_{j=1}^{n}P(c_{t+1}^j|c_t^j,w)} + \sum_{w}P(w)\,\log\frac{P(w)}{P(w)}$$

$$= \sum_{b,w} P(b,w)\,\log\frac{\prod_{j=1}^{n}P(c_{t+1}^j|c_t,w)}{\prod_{j=1}^{n}P(c_{t+1}^j|c_t^j,w)}.$$

The inequality above holds because the cross entropy is greater or equal to the entropy, see the relationship in (2.2). Therefore the new Integrated Information measure, the ground truth Integrated Information Φ_T, leads to the following expression

$$\inf_{Q\in\mathcal{M}_T} D_{\mathcal{B}\times\mathcal{W}_C}(P\parallel Q) = \sum_{b,w} P(b,w)\,\log\frac{\prod_{j=1}^{n}P(c_{t+1}^j|c_t,w)}{\prod_{j=1}^{n}P(c_{t+1}^j|c_t^j,w)}.$$

This can be rewritten as

$$\begin{aligned}
\inf_{Q\in\mathcal{M}_T} D_{\mathcal{B}\times\mathcal{W}_C}(P \parallel Q) &= \sum_{b,w} P(b,w)\, \log \frac{\prod_{j=1}^{n} P(c_{t+1}^j|c_t,w)}{\prod_{j=1}^{n} P(c_{t+1}^j|c_t^j,w)} \\
&= \sum_{b,w} P(b,w)\, \log \frac{\prod_{j=1}^{n} P(c_{t+1}^j,c_t,w)P(c_t^j,w)}{\prod_{j=1}^{n} P(c_{t+1}^j,c_t^j,w)P(c_t,w)} \\
&= \sum_{b,w} P(b,w)\, \log \frac{\prod_{j=1}^{n} P(c_{t+1}^j,c_t^{J\setminus\{j\}}|c_t^j,w)}{\prod_{j=1}^{n} P(c_{t+1}^j|c_t^j,w)P(c_t^{J\setminus\{j\}}|c_t^j,w)} \\
&= \sum_j I(C_{t+1}^j; C_t^{J\setminus\{j\}}|C_t^j, W_C).
\end{aligned}$$

□

The term $I(C_{t+1}^j; C_t^{J\setminus\{j\}}|C_t^j, W_C)$ is the conditional mutual information as defined in Definition 3. It characterizes the reduction of uncertainty in C_{t+1}^j due to $C_t^{J\setminus\{j\}}$ when W_C and C_t^j are given. Therefore this measure decomposes to a sum in which each addend characterizes the information flow towards one C_{t+1}^j. In terms of conditional independence statements Φ_T is 0 if and only if

$$C_{t+1}^j \perp\!\!\!\perp C_t^{J\setminus\{j\}} | \{C_t^j, W_C\}, \quad \text{for all } j \in J.$$

Ignoring W_C would lead exactly to the conditional independence statements in (3.3).

Since we have a graphical representation of our split model in this case, we know that $\Phi_T = 0$ if and only if the initial distribution $\hat{P}$ factors according to the graph that belongs to $\mathcal{M}_T$. This follows from Proposition 3 and the fact that the KL-divergence is 0 if and only if both distributions are equal. Hence we see once more that this measure truly removes the causal cross-connections. In the remainder of this section we explore different ways to write Φ_T and the corresponding conditional independence statements.

The exterior influence W_C is independent of C_t, $W_C \perp\!\!\!\perp C_t$, and therefore we are able to split up the conditional mutual information into a part corresponding to the conditional independence statements of Postulate 1 and another conditional mutual information:

$$\begin{aligned}
\Phi_T &= \sum_{j=1}^{n} I(C_{t+1}^j; C_t^{J\setminus\{j\}}|C_t^j, W_C) \\
&= \sum_{j=1}^{n} \sum_{c_{t+1}^j, c_t, w} P(b,w)\, \log \left(\frac{P(c_{t+1}^j, c_t^{J\setminus\{j\}}|c_t^j)}{P(c_{t+1}^j|c_t^j)P(c_t^{J\setminus\{j\}}|c_t^j)} \frac{P(c_{t+1}^j,c_t^j)P(c_t)P(c_{t+1}^j,c_t,w)P(c_t^j,w)}{P(c_{t+1}^j,c_t)P(c_t^j)P(c_{t+1}^j,c_t^j,w)P(c_t,w)} \right) \\
&= \sum_{j=1}^{n} I(C_{t+1}^j; C_t^{J\setminus\{j\}}|C_t^j) + \sum_{c_{t+1}^j, c_t, w} P(b,w)\, \log \frac{P(c_{t+1}^j,c_t^j)P(c_t)P(c_{t+1}^j,c_t,w)P(c_t^j,w)}{P(c_{t+1}^j,c_t)P(c_t^j)P(c_{t+1}^j,c_t^j,w)P(c_t,w)}.
\end{aligned}$$

Simple operations lead to

$$\Phi_T = \sum_{j=1}^{n} I(C_{t+1}^j; C_t^{J\setminus\{j\}} | C_t^j) + \sum_{c_{t+1}^j, c_t, w} P(b,w) \log \frac{P(w, c_t^{J\setminus\{j\}} | c_{t+1}^j, c_t^j)}{P(w|c_{t+1}^j, c_t^j) P(c_t^{J\setminus\{j\}} | c_{t+1}^j, c_t^j)}$$

$$= \sum_{j=1}^{n} I(C_{t+1}^j; C_t^{J\setminus\{j\}} | C_t^j) + I(C_t^{J\setminus\{j\}}; W_C | C_{t+1}^j, C_t^j).$$

Since the conditional mutual information is non-negative, Φ_T is 0 if and only if the CIS

$$C_{t+1}^j \perp\!\!\!\perp C_t^{J\setminus\{j\}} | C_t^j \quad \text{and} \quad C_t^{J\setminus\{j\}} \perp\!\!\!\perp W_C | \{C_t^j, C_{t+1}^j\}$$

hold for every $j \in J$. The first type of CISs are exactly the conditions for defining the set of split distributions in the context of the CIS Integrated Information, $\mathcal{M}_{CIS}$, in equation (3.3) and therefore the distributions in $\mathcal{M}_T$ satisfy Postulate 1.

Now we take a closer look at the second type of conditional independence statements. These are satisfied if and only if there is no reduction of uncertainty in $C_t^{J\setminus\{j\}}$ due to W_C given C_{t+1}^j, C_t^j. Although $C_t^{J\setminus\{j\}} \perp\!\!\!\perp W_C$ holds these variables are no longer independent once we condition on $\{C_{t+1}^j, C_t^j\}$. The reason for this is that C_{t+1}^j is a common effect of $C_t^{J\setminus\{j\}}$ and W_C.

Furthermore, we can show the following equivalencies between different formulations of the conditional independence statements.

Proposition 4. *Let $P \in \mathcal{M}_T^f$, then $W_C \perp\!\!\!\perp C_t$ and the following conditional independence statements are equivalent*

(1) $C_{t+1}^j \perp\!\!\!\perp C_t^{J\setminus\{j\}} | \{C_t^j, W_C\}$

(2) $C_{t+1}^j \perp\!\!\!\perp C_t^{J\setminus\{j\}} | C_t^j$ *and* $C_t^{J\setminus\{j\}} \perp\!\!\!\perp W_C | \{C_t^j, C_{t+1}^j\}$

(3) $C_t^{J\setminus\{j\}} \perp\!\!\!\perp \{W_C, C_{t+1}^j\} | C_t^j$

for $j \in J$.

Proof. The proof that (1) $\Rightarrow$ (2) is given above by

$$I(C_{t+1}^j; C_t^{J\setminus\{j\}} | C_t^j, W_C) = I(C_{t+1}^j; C_t^{J\setminus\{j\}} | C_t^j) + I(C_t^{J\setminus\{j\}}; W_C | C_{t+1}^j, C_t^j).$$

Now we show that (3) follows from (2). Applying the second and then the first CIS listed in (2) leads to the following equalities

$$P(c_t^{J\setminus\{j\}} | w, c_{t+1}^j, c_t^j) = P(c_t^{J\setminus\{j\}} | c_{t+1}^j, c_t^j) = P(c_t^{J\setminus\{j\}} | c_t^j)$$

for all $(b,w) \in \mathcal{B} \times \mathcal{W}_C$ and therefore (2) is equivalent to the CISs in (3).

It remains to show that (3) $\Rightarrow$ (1). Considering the CISs in (3) and that W_C is independent of C_t the equalities

$$\begin{aligned} P(c_t^{J\setminus\{j\}} | w, c_{t+1}^j, c_t^j) &= P(c_t^{J\setminus\{j\}} | c_t^j) \\ &= \frac{P(c_t)}{P(c_t^j)} \cdot \frac{P(w)P(c_t, w)}{P(w)P(c_t, w)} \\ &= P(c_t^{J\setminus\{j\}} | c_t^j, w) \frac{P(w)P(c_t)}{P(w, c_t)} \\ &= P(c_t^{J\setminus\{j\}} | c_t^j, w) \end{aligned}$$

hold for all $(b,w) \in \mathcal{B} \times \mathcal{W}_C$.

In the second line of the preceding equations we multiply the fraction by one and then use that $P(c_t, w)/(P(c_t^j)P(w)) = P(c_t^{J\setminus\{j\}}|c_t^j, w)$ to gain the third line. □

Note that in the calculations above we were able to utilize that we know the factorization of $\hat{P}$. Similar to the case of stochastic interaction, Φ_{SI}, we are also able to project a more general initial distribution $\hat{P} \in \mathcal{P}(\mathcal{B} \times \mathcal{W}_C)$ to $\mathcal{M}_T$. This leads to the following measure.

Proposition 5. *Let $\hat{P} \in \mathcal{P}(\mathcal{B} \times \mathcal{W}_C)$. Minimizing the KL-divergence between $\hat{P}$ and $\mathcal{M}_T$ leads to*

$$\begin{aligned}\inf_{Q\in\mathcal{M}_T} D_{\mathcal{B}\times\mathcal{W}_C}(\hat{P} \parallel Q) &= \sum_{b,w} \hat{P}(b,w) \log \frac{\hat{P}(b,w)}{\hat{P}(c_t)\hat{P}(w)\prod_{j=1}^{n} \hat{P}(c_{t+1}^j|c_t^j, w)} \\ &= H_{\hat{P}}(C_t) - H_{\hat{P}}(B|W_C) + \sum_j H_{\hat{P}}(C_{t+1}^j|C_t^j, W_C).\end{aligned}$$

The proof is analogous to the previous proofs where we use the fact that the cross entropy is greater or equal to the entropy.

Although this measure calculates the true value of Integrated Information, in general we do not necessarily know what the exterior influences on our system is. In these cases we have to assume that W_C is a hidden variable. This leads us to the measure Φ_{CII} that we discuss in the next section.

3.4 Causal Information Integration

In the previous section we defined the measure Φ_T under the assumption that all the influences on C_{t+1} are known. In many cases, however, there exist unknown influences on C_{t+1}. Hence we include W_C as a latent variable in order to model these unknown influences.

The necessity to consider an unknown influence was also discussed earlier in context of Postulate 2. We now utilize the notion of a common exterior influence to define the measure Φ_{CII}, which we call Causal Information Integration (CII). This measure should be used in case of an unknown exterior influence. Explicitly including said exterior influence allows us to avoid the problems of a fixed edge between the C_{t+1}^js. This leads to the graphical representation of the split models for $n = 2$ and $n = 3$ in Figure 3.7.

The factorization of the distributions belonging to these graphical models is the following one

$$P(b, w) = P(c_t) \prod_{j=1}^{n} P(c_{t+1}^j|c_t^j, w)P(w).$$

Note that this is the same factorization as in $\mathcal{M}_T$ in the previous section. Since W_C is a latent variable here, we marginalize over the elements of $\mathcal{W}_C$ and gain a distribution on $\mathcal{B}$ defining our new model instead of using the above factorization directly as split model. Additionally, the size of the state space $|\mathcal{W}_C| = m$ is unknown.

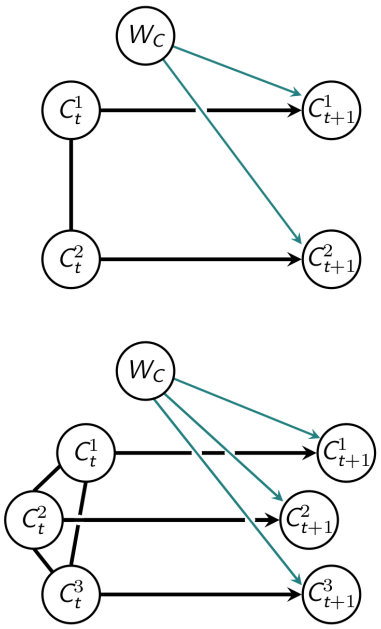


Figure 3.7: Split systems with exterior influences for $n = 2$ and $n = 3$.

Definition 18 (Causal Information Integration)**.** The set of distributions belonging to the marginalized model for $|\mathcal{W}_C^m| = m$ is

$$\mathcal{M}_{CII}^m = \left\{ P \in \mathcal{P}^\circ(\mathcal{B}) \,\middle|\, \exists Q \in \mathcal{P}^\circ(\mathcal{B} \times \mathcal{W}_C^m) : \right.$$
$$\left. P(b) = \sum_w Q(c_t)Q(w) \prod_{j=1}^n Q(c_{t+1}^j | c_t^j, w), \forall b \in \mathcal{B} \right\}.$$

We define the split model for this measure as the closure, denoted by a bar, of the union of $\mathcal{M}_{CII}^m$s:

$$\boxed{\mathcal{M}_{CII}} = \overline{\bigcup_{m \in \mathbb{N}} \mathcal{M}_{CII}^m}. \tag{3.5}$$

This leads to the measure

$$\boxed{\Phi_{CII}} = \inf_{Q \in \mathcal{M}_{CII}} D_{\mathcal{B}}(\hat{P} \parallel Q).$$

Since the split system $\mathcal{M}_{CII}$ was defined by marginalizing over elements in a graphical model, we are able to use the graphical representation to get a more precise notion of the cases in which $\Phi_{CII} = 0$ holds. In those cases the initial distribution can be completely explained as a limit of marginalized distributions without causal cross-connections and with exterior influences.

Proposition 6. *The measure Φ_{CII} is 0 if and only if there exists a sequence of distributions $Q^m \in \mathcal{P}(\mathcal{B})$ with the following properties.*

1. $\hat{P} = \lim_{m \to \infty} Q^m$.
2. *For every $m \in \mathbb{N}$ there exists a distribution $\hat{Q}^m \in \mathcal{P}(\mathcal{B} \times \mathcal{W}_C^m)$ that has $\mathcal{B}$ marginals equal to Q^m*

   $$Q^m(b) = \hat{Q}^m(b), \quad \forall b \in \mathcal{B}.$$

   Additionally $\hat{Q}^m$ factors according to the graph corresponding to the split system

   $$\hat{Q}^m(b, w) = \hat{Q}^m(c_t) \prod_{j=1}^n \hat{Q}^m(c_{t+1}^j | c_t^j, w) \hat{Q}^m(w), \quad \forall (b, w) \in \mathcal{B} \times \mathcal{W}^m.$$

Proof of Proposition 6. If $\Phi_{CII} = 0$ holds, then

$$\inf_{Q \in \mathcal{M}_{CII}} D_{\mathcal{B}}(\hat{P} \parallel Q) = 0.$$

Since $\mathcal{M}_{CII}$ is compact, the infimum is an element of $\mathcal{M}_{CII}$ so there exists $Q \in \mathcal{M}_{CII}$ such that $D_{\mathcal{B}}(\hat{P} \parallel Q) = 0$. Therefore $\hat{P} \in \mathcal{M}_{CII}$ and the existence of a sequence Q^m follows from the definition of $\mathcal{M}_{CII}$.

Assume that there exists a sequence Q^m that satisfies 1. and 2. Then every element $Q^m \in \mathcal{M}_{CII}^m$ per definition and the limit

$$\hat{P} \in \overline{\bigcup_{m \in \mathbb{N}} \mathcal{M}_{CII}^m} = \mathcal{M}_{CII}.$$

Hence

$$\Phi_{CII} = \inf_{Q \in \mathcal{M}_{CII}} D_{\mathcal{B}}(\hat{P} \parallel Q) = D_{\mathcal{B}}(\hat{P}, \hat{P}) = 0.$$

□

The set $\mathcal{M}_{CII}$ results from marginalizing over an element of a chain graph. In general this does not lead to a model corresponding to a new chain graph.

The Algorithm 11 from Reference [Sadeghi16], described in Section 2.4, allows us to transform a chain graph with latent variables into a chain mixed graph that represents the conditional independence structures of the marginalized chain graph. Using this on the graphs of the split systems in Figure 3.7 leads to the CMGs in Figure 3.8. In addition to the directed and undirected edges belonging to chain graphs the chain mixed graphs also have arcs $\leftrightarrow$. Two nodes connected by an arc are called spouses. The connection between spouses appears when we marginalize over a common influence, hence, spouses do not have a directed information flow from one node to the other but are affected by the same mechanisms. Unfortunately, there exists no new explicit factorization corresponding to the CMGs.

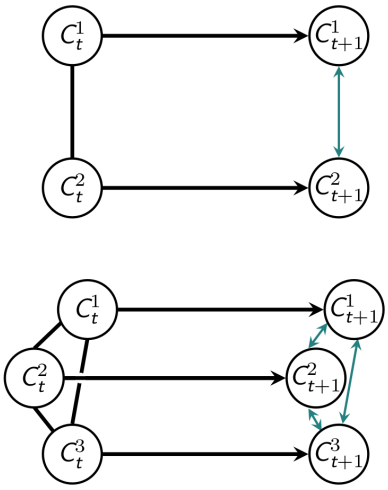


Figure 3.8: Marginalized model for $n = 2$ and 3.

The measure Φ_{CII} meets the requirements of both Postulate 1 and Postulate 2. In order to show that Φ_{CII} satisfies the conditional independence statements in Postulate 1 we calculate the conditional distributions $P(C_{t+1}^j|C_t^j)$ and $P(C_{t+1}^j|C_t)$ with respect to the factorization

$$P(b) = \sum_w P(c_t) \prod_{j=1}^{n} P(c_{t+1}^j|c_t^j, w)P(w), \quad \forall b \in \mathcal{B}.$$

This results in

$$\begin{aligned} P(c_{t+1}^j|c_t^j) &= \frac{\sum_{c_{t+1}^{J\setminus\{j\}}} \sum_{c_t^{J\setminus\{j\}}} \sum_w P(c_t) \prod_{j'=1}^{n} P(c_{t+1}^{j'}|c_t^{j'}, w)P(w)}{P(c_t^j)} \\ &= \frac{\sum_{c_t^{J\setminus\{j\}}} \sum_w P(c_t)P(c_{t+1}^j|c_t^j, w)P(w)}{P(c_t^j)} \\ &= \sum_w P(c_{t+1}^j|c_t^j, w)P(w) \end{aligned}$$

and

$$\begin{aligned} P(c_{t+1}^j|c_t) &= \frac{\sum_{c_{t+1}^{J\setminus\{j\}}} \sum_w P(c_t) \prod_{j'=1}^{n} P(c_{t+1}^{j'}|c_t^{j'}, w)P(w)}{P(c_t)} \\ &= \sum_w P(c_{t+1}^j|c_t^j, w)P(w) \end{aligned}$$

for all $b \in \mathcal{B}$. Hence, $P(C_{t+1}^j|C_t^j) = P(C_{t+1}^j|C_t)$ for every $P \in \mathcal{M}_{CII}^m$, $m \in \mathbb{N}$. Since every element $P \in \mathcal{M}_{CII}$ is a limit point of distributions that satisfy the conditional independence statements, P also fulfills those, as proven in Proposition 3.12 in [Lauritzen96]. Therefore, Φ_{CII} satisfies Postulate 1 and the set of all such distributions is a subset of $\mathcal{M}_{CIS}$

$$\mathcal{M}_{CII} \subseteq \mathcal{M}_{CIS}.$$

In order to prove that Φ_{CII} satisfies Postulate 2 we show that $\mathcal{M}_I$ is a subset of $\mathcal{M}_{CII}$ and therefore satisfies Postulate 3. At first we consider the following subset of $\mathcal{M}_{CII}$

$$\mathcal{M}_{CI}^m = \left\{ P \in \mathcal{P}(\mathcal{B}) \,\middle|\, \exists Q \in \mathcal{P}(\mathcal{B} \times \mathcal{W}_C^m) : P(b) = \sum_w Q(c_t)Q(w) \prod_{j=1}^n Q(c_{t+1}^j|w) \right\}$$

$$\mathcal{M}_{CI} = \overline{\bigcup_{m \in \mathbb{N}} \mathcal{M}_{CI}^m},$$

where we remove the connections between the different points in time, as shown in Figure 3.9.

Now C_t and C_{t+1} are independent of each other

$$Q(b) = Q(c_t)Q(c_{t+1})$$

with

$$Q(c_{t+1}) = \sum_w Q(w) \prod_{j=1}^n Q(c_{t+1}^j|w)$$

for $b \in \mathcal{B}$ and $Q \in \mathcal{M}_{CI}^m$. Since independence structures of discrete distributions are preserved in the limit, we have $\mathcal{M}_{CI} \subseteq \mathcal{M}_I$.

In order to gain equality it remains to show that $Q(C_{t+1})$ can approximate every distribution on $\mathcal{C}$ if the state space of W is sufficiently large. These distributions are mixtures of discrete product distributions, where

$$\prod_{j=1}^n Q(c_{t+1}^j|w)$$

are the mixture components and $Q(w)$ are the mixture weights. Hence, we are able to use the following result.

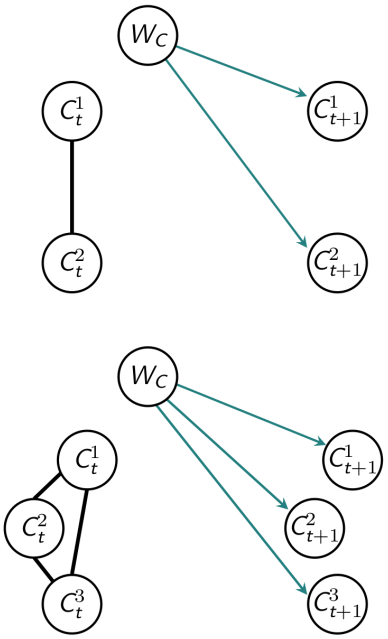


Figure 3.9: Submodels of the split models with exterior influences for $n = 2$ and $n = 3$.

Theorem 3.4.1 (Theorem 1.3.1 from Reference [Montúfar12]). *Let q be a prime power. The smallest m for which any probability distribution on $\{1, \ldots, q\}^n$ can be approximated arbitrarily well as mixture of m product distributions is q^{n-1}.*

The number q corresponds to the number of states of one variable C_{t+1}^j and n is the number of variables in one point in time. Hence, if we have, for example, three binary variables, then the theorem above states that any distribution on $\mathcal{C}$ can be approximated arbitrarily well if the state space of W_C has at least $2^{(3-1)} = 4$ states. In conclusion, $\mathcal{M}_I \subseteq \mathcal{M}_{CI} \subseteq \mathcal{M}_{CII}$ and so Φ_{CII} satisfies Postulate 2 in addition to Postulate 1.

Universal approximation results like the previous theorem might suggest that the model $\mathcal{M}_{CII}$ is expressive enough to approximate a model with an undirected edge between the C_{t+1}^js. This cannot be the case because then $\mathcal{M}_{CII}$ would include the split models of the geometric Integrated Information Φ_G discussed in Section 3.2.2. This is not possible because all the elements in the closed set $\mathcal{M}_{CII}$ satisfy the conditional independence statements of Postulate 1, but the distributions in $\mathcal{M}_G$ do not satisfy these conditions in general. We discuss the relationships between the different measures in more detail in Section 3.5.

Note that using Φ_{CII} in cases without or with a very small exterior influence might not capture all the causal cross-connections, since the additional latent variable can compensate

for some of the causal cross-connections. This can only be avoided when the exterior influence is known and can therefore be included in the model explicitly. In this case the measure would result in the ground truth Integrated Information Φ_T from the previous section.

Now the question naturally arises whether we are able to include further exterior influences on our model in order to gain more general distributions and a split model larger than $\mathcal{M}_{CII}$. This would then lead to the definition of a new Integrated Information measures bound from above by the causal information integration. We explore this thought by starting with the graph corresponding to a model without any exterior influences. This is the split model $\mathcal{M}_{SI}$ associated to the stochastic interaction Φ_{SI} and depicted in Figure 3.10 on the left. Then we add vertices corresponding to additional latent variables and the respective edges to the graph in a way such that the whole graph is still a chain graph. An example of this procedure is depicted in Figure 3.10 in the middle.

Since we are interested in the connections among the visible variables after marginalization, it is only important how the visible nodes are connected via the hidden nodes. In the case of the example in Figure 3.10 this leads to a directed path from C_t^1 to C_t^2 going through the hidden nodes. Therefore we are able to reduce the hidden structure to a gray box shown on the right in Figure 3.10.

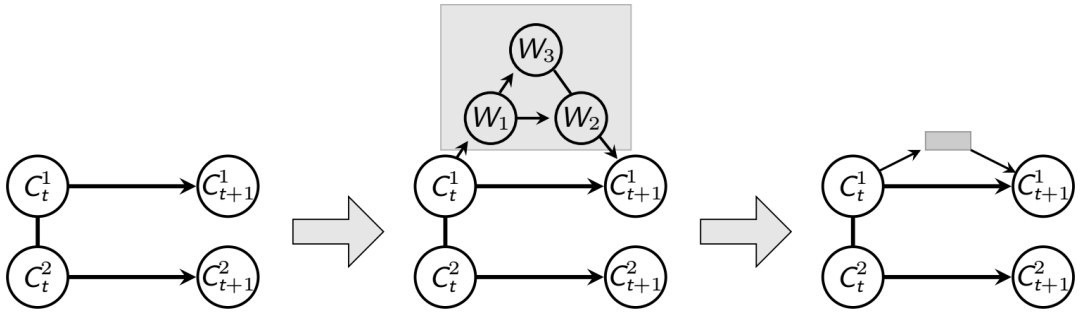


Figure 3.10: The graph corresponding to $\mathcal{M}_{SI}$ on the left, with an exterior influence in the middle and with the reduced exterior influence on the right.

Once we have this type of graph we can use the Algorithm 11, given in Section 2.4.4, which converts a chain graph with latent variables to a chain mixed graph reflecting the conditional independence structure of the marginalized model. In our example this leads to a directed edge from C_t^1 to C_t^2. Seeing that this directed edge already existed the resulting model now is a subset of $\mathcal{M}_{SI}$ and is therefore a proper subset of $\mathcal{M}_{CII}$.

Following this line of reasoning we are able to prove that adding further hidden nodes and subgraphs of hidden nodes does not lead to a chain mixed graph belonging to a model that satisfies the CIS of Postulate 1 and strictly contains $\mathcal{M}_{CII}$.

Theorem 3.4.2. *It is not possible to create a chain mixed graph corresponding to a model $\mathcal{M}$, such that its distributions satisfy Postulate 1 and $\mathcal{M}_{CII} \subsetneq \mathcal{M}$ by introducing a more complicated hidden structure to the graph of $\mathcal{M}_{SI}$.*

Proof of Theorem 3.4.2. At first we consider the case of two nodes per time step, $n = 2$. We now take a closer look at the possible ways a hidden structure could be connected to the left graph in Figure 3.11. There we discuss all the possible connections that exist between two nodes, as depicted on the right in Figure 3.11. The boxes represent any kind of subgraph of hidden nodes, as discussed above, such that the whole graph is still a chain graph and the two headed dotted arrows stand for an undirected edge or an arrow in any direction.

Consider two nodes C^1 and C^2 then the connections including a box between the nodes can take one of the five following forms

(1) they form an undirected path between C^1 and C^2,
(2) they form a directed path from C^1 to C^2,
(3) they form a directed path from C^2 to C^1,
(4) there exists a collider or
(5) C^1 and C^2 have a common exterior influence.

A collider is a node or a set of nodes connected by undirected edges that have an arrow pointing at the set at both ends, more precisely it has the form $\rightarrow \bullet \cdots \bullet \leftarrow$.

We will start with the gridded hidden structure connected to C_t^1 and C_t^2 in Figure 3.11. Since there already is an undirected edge between the C_t^js, an undirected path would make no difference for the marginalized model. The cases (2) and (3) would form a semi-directed cycle which violates the requirements of a chain mixed graph. A collider would also make no difference since it disappears in the marginalized model.

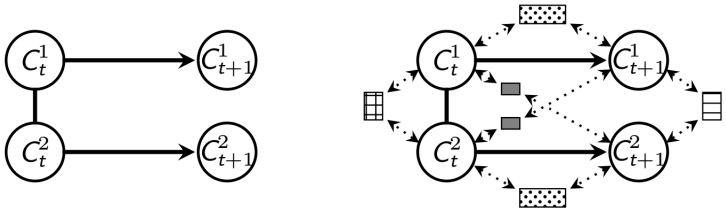


Figure 3.11: The graph corresponding to $\mathcal{M}_{SI}$ on the left and possible connections between two vertices in this graph and a hidden structure, depicted on the right.

Furthermore, a common exterior influence would lead to

$$P(w)P(c_t|w)P(c_{t+1}^1|c_t^1)P(c_{t+1}^2|c_t^2) = P(c_t, w)P(c_{t+1}^1|c_t^1)P(c_{t+1}^2|c_t^2)$$
$$\sum_w P(c_t, w)P(c_{t+1}^1|c_t^1)P(c_{t+1}^2|c_t^2) = P(c_t)P(c_{t+1}^1|c_t^1)P(c_{t+1}^2|c_t^2).$$

Therefore a common exterior influence on the vertices in timestep t would not add any additional elements to $\mathcal{M}_{CII}$.

Now we discuss the possibilities in the case of a hidden structure between the nodes $C_t^{j'}$ and C_{t+1}^j, $j', j \in \{1, 2\}$, $j' \neq j$. An undirected edge or a directed edge (3) would create a directed cycle. A directed path (2) from $C_t^{j'}$ to C_{t+1}^j would lead to a chain graph in which $C_t^{j'}$ and C_{t+1}^j are not conditionally independent given C_t^j, hence it would not satisfy Postulate 1. If there exists a collider (4) in the hidden structure, then nothing else in the graph depends on this part of the structure and it reduces to a factor one when we marginalize over the hidden variables. Therefore the path between $C_t^{j'}$ and C_{t+1}^j gets interrupted leaving only a potential external influence. A common exterior influence (5) leads to a chain mixed graph that does not satisfy the necessary conditional independence statements, because using the Algorithm 11 leads to an arc between $C_t^{j'}$ and C_{t+1}^j. Hence, they are c-connected in the sense of Definition 10.

The next possibility is a dotted hidden structure between C_t^j and C_{t+1}^j, $j \in \{1, 2\}$. An undirected path (1) and a directed path (3) would lead to a semi-directed cycle. A directed path (2) would add no new structure to the model since there already is a directed edge

between C_t^j and C_{t+1}^j. A collider (4) does not have an effect on the marginalized model. Adding a common exterior influence W_1 on C_t^1, C_{t+1}^1 results in a new model that is not symmetric in $j \in \{1,2\}$ and does not include $\mathcal{M}_I$, therefore it does not fully contain $\mathcal{M}_{CII}$. Adding additional common exterior W_2 influences on C_t^2, C_{t+1}^2 or C_{t+1}^1, C_{t+1}^2, in order to include $\mathcal{M}_I$ in the new model, violates the CIS of Postulate 1 since the nodes in W_1 and W_2 are connected in the moralized graph.

The last hidden structure between two nodes is the striped one that connects the C_{t+1}^js. An undirected path (1) or any directed path (2),(3) lead to a graph that does not satisfy the necessary CIS. A collider (4) has no impact on the model and a common exterior influence leads exactly to the definition of the split system $\mathcal{M}_{CII}$.

Therefore, we are now considering connections among a hidden structure and three different visible nodes. Connecting C_{t+1}^1, C_{t+1}^2 and $C_t^j, j \in \{1,2\}$ leads either to a violation of the conditional independence statements or creates a semi-directed or directed cycle.

All the possible ways a hidden structure could be connected to three nodes C_t^1, C_t^2, C_{t+1}^1 by directed edges are shown in Figure 3.12. Replacing any of these edges by an undirected edge would either make no difference or lead to a model that does not satisfy the conditional independence statements. Here the boxes represent sections since more complicated hidden structures reduce to this case after marginalization.

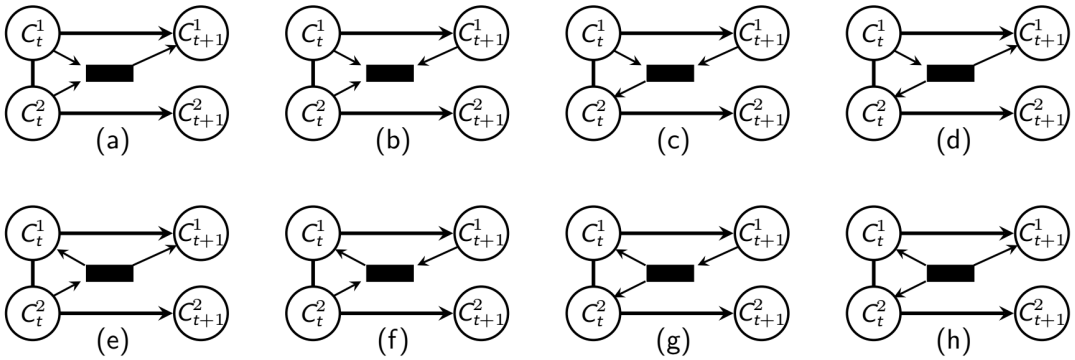


Figure 3.12: The eight possible hidden structures between the three vertices C_t^1, C_t^2, and C_{t+1}^1.

The models in (c), (d), (e), (f) and (g) each induce a directed cycle. We see that (a) and (h) display structures that do not satisfy the required CIS since the hidden structure establishes a connection between C_t^2 and C_{t+1}^1. The hidden structure in (b) has no impact on the marginalized model.

A hidden structure connected to all four nodes contains one of the structures above and therefore does not induce a new valid model.

Let us now consider a model with $n > 2$. Any hidden structure on this model either connects only up to four nodes and reduces therefore to one of the cases above, contains one of the connections discussed in Figure 3.12 or only connects nodes among one point in time. The only additional structures that we are able to consider to add would be a common exterior influence on the C_t^js, a common exterior influence on the C_{t+1}^js or a collider section on any nodes. All these structures do not change the marginalized model and therefore they would not lead to a set that strictly contains $\mathcal{M}_{CII}$.

Therefore it is not possible to create a chain graph with hidden nodes in order to get a model larger than $\mathcal{M}_{CII}$. □

3.4.1 Calculation of the Causal Information Integration

In this section we discuss how to calculate the measure Φ^m_{CII} with

$$\Phi^m_{CII} = \inf_{Q \in \mathcal{M}^m_{CII}} D_{\mathcal{B}}(\hat{P} \parallel Q)$$

by using the *em*-algorithm. This algorithm iterates between performing an m- and an e-projection between two sets in order to find the minimum KL-divergence between them, as defined in more detail in Section 2.5.1.

In order to calculate the KL-divergence between the distribution $\hat{P}$ and the set $\mathcal{M}^m_{CII}$ on $\mathcal{B}$ we consider the extended space of distributions on $\mathcal{B} \times \mathcal{W}^m_C$ denoted by $\mathcal{P}(\mathcal{B} \times \mathcal{W}^m_C)$. Let $\mathcal{M}_{W|B}$ be defined as the set of all distributions on $\mathcal{B} \times \mathcal{W}^m_C$ that have $\mathcal{B}$-marginals equal to the distribution of the whole system $\hat{P}$

$$\begin{aligned}\mathcal{M}_{W|B} &= \left\{ P \in \mathcal{P}^\circ(\mathcal{B} \times \mathcal{W}^m_C) \mid P(b) = \hat{P}(b),\ \forall b \in \mathcal{B} \right\} \\ &= \left\{ P \in \mathcal{P}^\circ(\mathcal{B} \times \mathcal{W}^m_C) \mid P(b,w) = \hat{P}(b)P(w|b),\ \forall (b,w) \in \mathcal{B} \times \mathcal{W}^m_C \right\}.\end{aligned}$$

We show below that there exists a unique e-projection to $\mathcal{M}_{W|B}$.

The second set is the set $\mathcal{E}^m$ of distributions that factor according to the split model including the common exterior influence. Note that if the size of the state space were known, then this set would correspond to the set of split systems $\mathcal{M}_T$ defined for the ground truth Integrated Information in Section 3.3.

$$\mathcal{E}^m = \left\{ P \in \mathcal{P}^\circ(\mathcal{B} \times \mathcal{W}^m_C) \,\middle|\, P(b,w) = P(c_t) \prod_{j=1}^{n} P(c^j_{t+1} | c^j_t, w) P(w),\ \forall (b,w) \in \mathcal{B} \times \mathcal{W}^m_C \right\} \tag{3.6}$$

Summing over the hidden variable W for an element in $\mathcal{E}^m$ would lead to an element in $\mathcal{M}^m_{CII}$. The set $\mathcal{M}^m_{CII}$ is a stratified exponential family, which is a finite union of curved exponential families satisfying some regularity conditions, as described in Section 2.3, and therefore has in general no unique m-projection. For the set $\mathcal{E}^m$, on the other hand, we show that there exists such a projection.

Now we show how to perform an e- and m-projection in this case. The e-projection from $Q \in \mathcal{E}^m$ to $\mathcal{M}_{W|B}$ is given by

$$P(b,w) = \hat{P}(b)Q(w|b),$$

for all $(b,w) \in \mathcal{B} \times \mathcal{W}^m_C$. This is the projection because of the following equality

$$\begin{aligned}D_{\mathcal{B} \times \mathcal{W}^m}(P \parallel Q) &= \sum_{b,w} P(b,w) \log \frac{P(b,w)}{Q(b,w)} \\ &= \sum_b \hat{P}(b) \log \frac{\hat{P}(b)}{Q(b)} + \sum_{b,w} P(b,w) \log \frac{P(w|b)}{Q(w|b)}.\end{aligned}$$

The first addend is a constant for a fixed distribution $\hat{P}$ and the second addend is equal to 0 if and only if $P(w|b) = Q(w|b)$.

After discussing the e-projection we now consider the m-projection.

Proposition 7. *The m-projection from $P \in \mathcal{M}_{W|B}$ to $\mathcal{E}^m$ is given by*

$$Q(b,w) = P(c_t) \prod_{j=1}^{n} P(c_{t+1}^j | c_t^j, w) P(w)$$

for all $(b,w) \in \mathcal{B} \times \mathcal{W}_C^m$.

Proof of Proposition 7. The KL-divergence between $P \in \mathcal{M}_{W|B}$ and $Q \in \mathcal{E}^m$ can be written as

$$\begin{aligned}
D_{\mathcal{B}\times\mathcal{W}_C^m}(P \parallel Q) &= \sum_{b,w} P(b,w) \log \frac{P(b,w)}{Q(c_t) \prod_{j=1}^{n} Q(c_{t+1}^j | c_t^j, w) Q(w)} \\
&= \sum_{b,w} P(b,w) \log P(b,w) \\
&\quad + \sum_{b,w} P(b,w) \log \frac{1}{Q(c_t)} \\
&\quad + \sum_{b,w} \sum_{j=1}^{n} P(b,w) \log \frac{1}{Q(c_{t+1}^j | c_t^j, w)} \\
&\quad + \sum_{b,w} P(b,w) \log \frac{1}{Q(w)}.
\end{aligned}$$

The first addend is a constant for P and the other addends are cross-entropies, which are greater or equal to the entropy. Therefore this leads to the following inequality.

$$\begin{aligned}
D_{\mathcal{B}\times\mathcal{W}_C^m}(P \parallel Q) &\geq \sum_{b,w} P(b,w) \log P(b,w) \\
&\quad + \sum_{b,w} P(b,w) \log \frac{1}{P(c_t)} \\
&\quad + \sum_{b,w} \sum_{j=1}^{n} P(b,w) \log \frac{1}{P(c_{t+1}^j | c_t^j, w)} \\
&\quad + \sum_{b,w} P(b,w) \log \frac{1}{P(w)} \\
&= \sum_{b,w} P(b,w) \log \frac{P(b,w)}{P(c_t) \prod_{j=1}^{n} P(c_{t+1}^j | c_t^j, w) P(w)}.
\end{aligned}$$

Hence, this projection is unique. □

Performing these projections iteratively leads to the minimization of the KL-divergence between $\mathcal{M}_{W|B}$ and $\mathcal{E}^m$, which is equivalent to an m-projection of $\hat{P}$ to $\mathcal{M}_{CII}^m$

$$\inf_{P \in \mathcal{M}_{W|B}, Q \in \mathcal{E}^m} D_{\mathcal{B}\times\mathcal{W}}(P \parallel Q) = \inf_{Q \in \mathcal{M}_{CII}^m} D_{\mathcal{B}}(\hat{P} \parallel Q).$$

This is the statement in Theorem 1.

Before we are able to calculate the causal information integration for a fixed m, Φ^m_{CII}, we still need to choose the initial distribution. Since the e- and m-projections are unique in our case, the outcome of the minimization depends only on the initial distribution. Thereby the initial distribution determines whether the algorithm converges to a local or global minimum. Hence, it is important to take the minimal outcome of multiple runs. One class of starting points that lead to a minimum that is in general not minimal is the one in which B and W are independent for all distributions $P^0(b,w) = P^0(b)P^0(w)$, for all $(b,w) \in \mathcal{B} \times \mathcal{W}$. It is easy to check that in this case the em-algorithm converges to the fixed point P' with

$$P'(b,w) = \hat{P}(c_t)\frac{1}{|\mathcal{W}_C^m|}\prod_{j=1}^{n}\hat{P}(c_{t+1}^j|c_t^j)$$

$$P'(b) = \hat{P}(c_t)\prod_{j=1}^{n}\hat{P}(c_{t+1}^j|c_t^j)$$

for all $(b,w) \in \mathcal{B} \times \mathcal{W}$.

Note, that this is the result of the m-projection of $\hat{P}$ to $\mathcal{M}_{SI}$. Hence, the algorithm converges to a local minimum that corresponds to the stochastic interaction, Φ_{SI}, if we start with an initial distribution in which $W \perp\!\!\!\perp B$.

Now we discuss the behavior of the different local minima in the next example.

Example 1. Here we are using a Boltzmann machine inspired approach as described in Section 2.1 with 5 controller nodes, no exterior influence and the matrix V with

$$V = \begin{pmatrix} -0.356 & -0.0978 & 0.897 & -0.006 & -0.039 \\ -0.226 & 0.478 & -0.430 & 0.187 & 0.251 \\ -0.861 & -0.183 & -0.715 & -0.081 & -0.644 \\ -0.14 & -0.032 & -0.811 & -0.333 & -0.574 \\ 0.189 & -0.990 & 0.321 & 0.691 & -0.692 \end{pmatrix}.$$

Using this example, we take a closer look at the local minima the em-algorithm converges to. Calculating Φ^m_{CII} for varying sizes of the state space, more precisely for $m = 2, 4, 8$ and 16, leads to the upper part in Figure 3.13.

This top half of the figure displays ten different runs of the em-algorithm with each size of state space in different shades of blue. The shade of blue darkens with the size of the state space. Here we display the outcomes of every run and not only the minimal one since we are interested in the local minima. We observe that increasing the state space leads to a smaller value of Φ^m_{CII}. Additionally, the differences between the minimal values corresponding to each state space grow smaller and converge as the size of the state spaces increases.

The bottom half of Figure 3.13 highlights the following observation. Each of the four illustrations is a copy of the one above where additionally the differences between the local minima for one m are shaded. By increasing the size of the state space the difference in value between the various local minima decreases visibly. This is consistent with the general observation made in the context of high dimensional optimization, for example, Reference [Choromanska15] in which the authors conjecture that the probability of finding a high valued local minimum decreases when the network size grows.

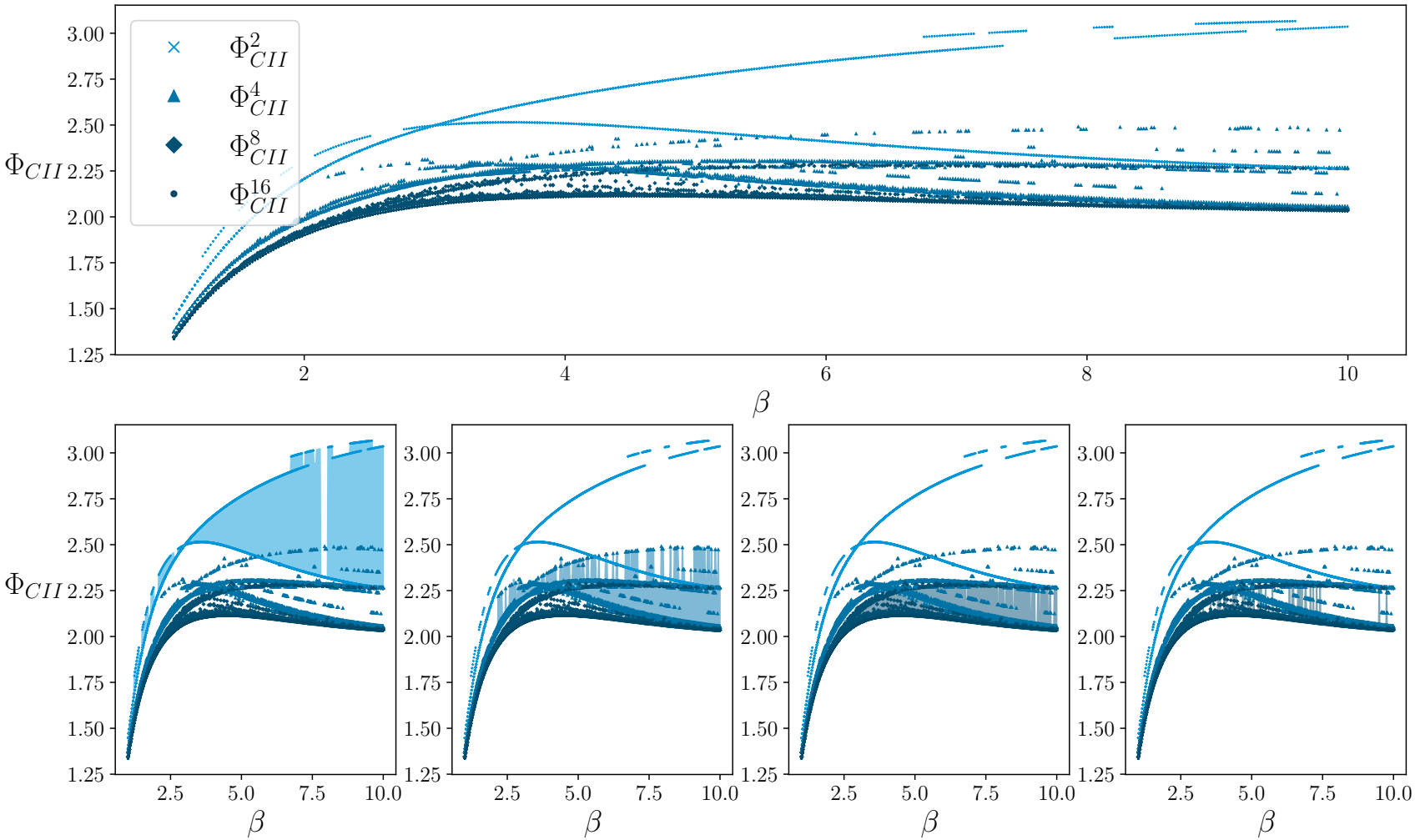


Figure 3.13: The outcome of 10 different initial distributions for Φ^m_{CII} with $m = 2, 4, 8$ and 16 on the top and for each of the sizes of m one figure on the bottom where the differences between the local minima are shaded.

Now we use the same example but we let the algorithm run only once with $|\mathcal{W}_C| = 2$. This leads to the results on the left in Figure 3.14.

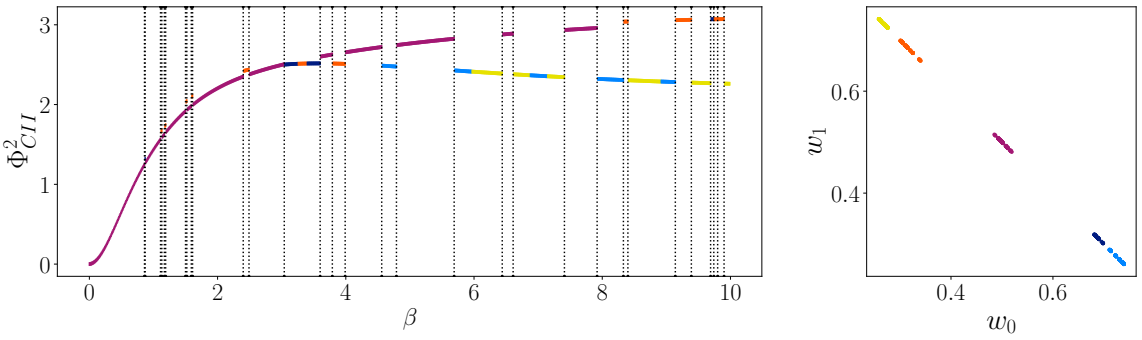


Figure 3.14: Results of one run of the *em*-algorithm for Φ^2_{CII} with each point colored according to the distribution of W_C, depicted on the right.

The interpretation of these results is the following one. The sets $\mathcal{E}^2$ and $\mathcal{M}^2_{CII}$, defined in the equations (3.6) and (3.5), do not change for different values of the inverse temperature β. Therefore we have a fixed set of local minima for a fixed size of the state space of W_C.

The element of this minimization that does change with different β is the initial distribution $\hat{P}$ and therefore which of the local minima are global minima.

The vertical dotted lines in Figure 3.14 represent the steps from P^{β_t} to $P^{\beta_{t+1}}$ in which the KL-divergence between the projection to $\mathcal{M}_{CII}$ is greater than 0.2

$$D_{\mathcal{B}}(P^{\beta_t,\star} \parallel P^{\beta_{t+1},\star}) > 0.2.$$

This means that inside the different sections of the curve the projections to $\mathcal{M}^2_{CII}$ are close to each other. As β increases a different region of local minima becomes global, which results in a new section in Figure 3.14. A sketch of this is depicted in Figure 3.15.

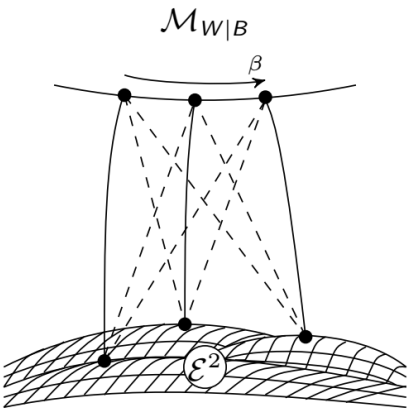


Figure 3.15: Sketch of different local minima as we increase β.

The curve in Figure 3.14 is colored according to the probability distribution of W_C as shown on the right side of Figure 3.14.
We see that a different distribution on $\mathcal{W}_C$ results in a different minimum except for the region around approx. 6, 7 and 8.5. There the regions are colored light blue and yellow. These colors refer to distributions on $\mathcal{W}_C$ that are different but symmetric in the following way. Consider two different distributions Q, Q' on $\mathcal{B} \times \mathcal{W}_C$ such that

$$Q(b, w_1) = Q'(b, w_2) \text{ and } Q(b, w_2) = Q'(b, w_1)$$

for all $b \in \mathcal{B}$ and $\mathcal{W}_C = \{w_1, w_2\}$. Then the corresponding marginalized distributions of Q and Q' in $\mathcal{M}^2_{CII}$ are equal

$$\sum_w Q(b, w) = \sum_w Q'(b, w_1).$$

This symmetry is the reason for the different colors in the blue and yellow regions.

By using the information geometric *em*-algorithm we therefore gain a notion of the local minima on $\mathcal{E}^2$.

3.5 Relationships between the different Integrated Information Measures

In this section we analyze the relationships among the different Integrated Information measures introduced in the previous sections, namely Φ_I, Φ_{SI}, Φ_G, Φ_{CIS}, Φ_T and Φ_{CII}. We first describe their relations in detail and then give a summary of them in the next section, Section 3.5.1.

We start with the measure Φ_I, which was discussed in relation to Postulate 2. Every measure that satisfies this postulate is bounded from above by the mutual information, that is by Φ_I. Hence, this results in the relationships

$$\begin{aligned}\Phi_I &\geq \Phi_G,\\ \Phi_I &\geq \Phi_{CIS}\\ \Phi_I &\geq \Phi_{CII}.\end{aligned}$$

In Example 3 in Section 3.5.1 we can see a situation in which Φ_{SI} and Φ_T exceed the value of Φ_I.

The split model for the stochastic interaction, $\mathcal{M}_{SI}$, is a subset of the split models for the measures geometric Integrated Information $\mathcal{M}_G$, CIS Integrated Information $\mathcal{M}_{CIS}$ and causal Integrated Information $\mathcal{M}_{CII}$. Therefore

$$\begin{aligned}\Phi_{SI} &\geq \Phi_G,\\ \Phi_{SI} &\geq \Phi_{CIS}\\ \Phi_{SI} &\geq \Phi_{CII}\end{aligned}$$

hold. Note that in the case in which the exterior influence has no impact on our system $\Phi_{SI} = \Phi_T$. In that case $\hat{P} \in \mathcal{P}(\mathcal{B} \times \mathcal{W}_C)$ with $W_C \perp\!\!\!\perp Y_t, Y_{t+1}$, which leads to the equality

$$\begin{aligned}\Phi_T &= \sum_{b,w} \hat{P}(b,w) \log \frac{\hat{P}(b)\hat{P}(w)}{\hat{P}(c_t)\hat{P}(w)\prod_{j=1}^{n}\hat{P}(c_{t+1}^j|c_t^j,w)}\\ &= \sum_{b,w} \hat{P}(b,w) \log \frac{\hat{P}(c_{t+1}|c_t)}{\prod_{j=1}^{n}\hat{P}(c_{t+1}^j|c_t^j)}\\ &= \Phi_{SI}.\end{aligned}$$

Next we look at the geometric Integrated Information measure, Φ_G. This measure is the only one of the discussed Integrated Information measures that does not satisfy Postulate 1. Therefore $\mathcal{M}_G \nsubseteq \mathcal{M}_{CIS}$ and $\mathcal{M}_G \nsubseteq \mathcal{M}_{CII}$. To evaluate the other inclusions we consider the more refined parametrizations of elements $P \in \mathcal{M}_{CII}^m$ and $Q \in \mathcal{M}_G$ as defined Section 2.4.3. These are

$$\begin{aligned}P(b) =& P(c_t)f_2(c_t^1, c_{t+1}^1)g_2(c_t^2, c_{t+1}^2)\\ &\cdot \sum_w P(w)f_1(w, c_{t+1}^1)f_3(c_t^1, c_{t+1}^1, w)g_1(w, c_{t+1}^2)g_3(c_t^2, c_{t+1}^2, w)\\ =& P(c_t)f_2(c_t^1, c_{t+1}^1)g_2(c_t^2, c_{t+1}^2)\phi(c_t^1, c_t^2, c_{t+1}^1, c_{t+1}^2),\\ Q(b) =& h_{n+1}(c_t)h_{n+2}(c_{t+1})\prod_{j=1}^{n} h_j(c_{t+1}^j, c_t^j),\end{aligned}$$

where $f_1, f_2, f_3, g_1, g_2, g_3, h_1, h_2, h_3, h_4$ are non-negative functions such that $P, Q \in \mathcal{P}(\mathcal{B})$ and

$$\phi(c_t^1, c_t^2, c_{t+1}^1, c_{t+1}^2) = \sum_w P(w) f_1(w, c_{t+1}^1) f_3(c_t^1, c_{t+1}^1, w) g_1(w, c_{t+1}^2) g_3(c_t^2, c_{t+1}^2, w).$$

Since ϕ depends on more than C_{t+1}^1 and C_{t+1}^2, $P(b)$ does not factorize according to $\mathcal{M}_G$ in general. Hence, $\mathcal{M}_{CII} \nsubseteq \mathcal{M}_G$ holds and since $\mathcal{M}_{CII} \subseteq \mathcal{M}_{CIS}$ we also have $\mathcal{M}_{CIS} \nsubseteq \mathcal{M}_G$. We observe in Example 3 that Φ_G can be greater or smaller than Φ_{CIS}, Φ_T or Φ_{CIS}.

Minimizing over all possible external influences W might compensate for a part of the causal cross-connection that we aim at measuring. One example in which accounting for an exterior influence that does not exist leads to a value smaller than the true Integrated Information is discussed earlier in the context of Postulate 2. There we refer to an example in [Kanwal17] where Φ_{SI} exceeds Φ_I in a setting without an exterior influence. Similarly, Φ_{CII} is smaller or equal to the true value Φ_T.

Furthermore, looking at the parametrizations allows us to identify a subset of distributions that lies in the intersection of $\mathcal{M}_G$ and $\mathcal{M}_{CII}$. Allowing P to only have pairwise interactions would lead to

$$\begin{aligned} P(b) &= P(c_t) \tilde{f}_2(c_t^1, c_{t+1}^1) \tilde{g}_2(c_t^2, c_{t+1}^2) \sum_w P(w) \tilde{f}_1(w, c_{t+1}^1) \tilde{g}_1(w, c_{t+1}^2) \\ &= P(c_t) \tilde{f}_2(c_t^1, c_t^1) \tilde{g}_2(c_t^2, c_{t+1}^2) \tilde{\phi}(c_{t+1}^1, c_{t+1}^2), \end{aligned}$$

with the non-negative functions $\tilde{f}_1, \tilde{f}_2, \tilde{g}_1, \tilde{g}_2$ such that $P \in \mathcal{P}(\mathcal{B})$ and

$$\tilde{\phi}(c_{t+1}^1, c_{t+1}^2) = \sum_w P(w) \tilde{f}_1(w, c_{t+1}^1) \tilde{g}_1(w, c_{t+1}^2).$$

This P is an element of $\mathcal{M}_G \cap \mathcal{M}_{CII}$.

In the next part we discuss the relationship between $\mathcal{M}_{CII}$ and $\mathcal{M}_{CIS}$. As stated before the elements in $\mathcal{M}_{CII}$ satisfy the conditional independence statements of Postulate 1 and therefore

$$\mathcal{M}_{CII} \subseteq \mathcal{M}_{CIS}$$

and

$$\Phi_{CIS} \leq \Phi_{CII}.$$

The question remains whether $\mathcal{M}_{CIS}$ is contained in $\mathcal{M}_{CII}$, which would lead to an equivalence between the two measures. Previously we have seen that making the state space of W_C large enough, in the case of no information flow between the points in time, can approximate any distribution between the C_{t+1}^js, see Theorem 3.4.1. This gives the impression that the models $\mathcal{M}_{CII}$ and $\mathcal{M}_{CIS}$ might actually be equal.

The lack of a graphical representation for Φ_{CIS} and of a closed form solution for both of them impede the analysis of their relationship. However, based on numerically calculated examples we have the following conjecture.

Conjecture 1. *It is not possible to approximate every distribution $Q \in \mathcal{M}_{CIS}$ with arbitrary accuracy by an element of $P \in \mathcal{M}_{CII}$. Therefore, the following relationship holds*

$$\mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS}.$$

The following example suggests that this conjecture might hold.

Example 2. Here, we use the Boltzmann machine inspired approach, discussed in Section 2.1, with 3 visible binary variables and no exterior influence. Hence, in this case the true measure for Integrated Information is given by the stochastic interaction, $\Phi_T = \Phi_{SI}$. The connection matrix V that leads to the results in Figure 3.16 is depicted below

$$V = \begin{pmatrix} -0.435 & 0.474 & 0.368 \\ 0.521 & 0.007 & -0.739 \\ -0.561 & -0.969 & -0.764 \end{pmatrix}.$$

Note that Φ_{CIS} was calculated numerically by using the minimize function from the scipy.optimize package in python. There we apply the "trust-constr" approach that uses a trust region method, which allows us to minimize a scalar function, in our case the KL-divergence, subject to the conditional independence constraints.

For each β we optimize with 100 different initial distributions and take the minimum of the results because this method is not guaranteed to result in a global minimum. This is why there are a few points, for example at $\beta = 1.9$, where the reached minimum might not be optimal. Even with this caveat there still is a visible difference between Φ^m_{CII} for $m = 4096$ and Φ_{CIS}.

Although this example shows that $\mathcal{M}^m_{CII} \subsetneq \mathcal{M}_{CIS}$ for $m = 4096$ it might be the case that this latent space is still not large enough in order to approximate $\mathcal{M}_{CIS}$ well. However, if we need more than 2^{12} states for three binary variables in each point in time, that have in total 64 different states, does this really describe a reasonable situation in which we want to calculate the Integrated Information? In the scenario in which we can explain almost every distribution on the C^j_{t+1}s via an exterior influence, why would we attribute any importance to the information flow inside the system in the first place? Additionally, increasing the state space of W_C much further can lead to an intractable problem. Therefore we would argue that it is more plausible to use Φ^m_{CII} with a reasonably sized m depending on the application.

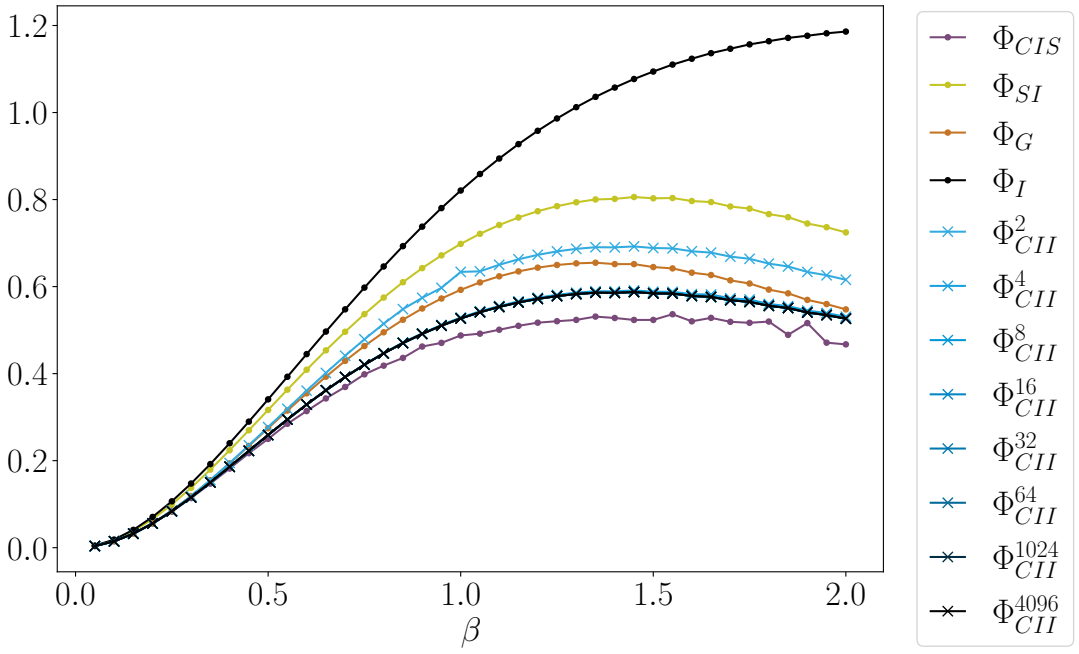


Figure 3.16: The results for the different Integrated Information measures in the case of $n = 3$ and no exterior influence.

Lastly, we explore the relationship between Φ_T and Φ_{CII} in detail.

Proposition 8. *Let $\hat{P} \in \mathcal{P}(\mathcal{B} \times \mathcal{W}_C)$. Then the ground truth Integrated Information, Φ_T, is an upper bound for Φ_{CII},*

$$\Phi_{CII} \leq \Phi_T.$$

Additionally, they are equal if and only if $\Phi_T = 0$.

Proof of Proposition 8. Let $m = |\mathcal{W}_C|$ the true size of the state space. By using the log-sum inequality we get

$$\begin{aligned}
\Phi^m_{CII} &= \inf_{Q \in \mathcal{M}^m_{CII}} \sum_b \hat{P}(b) \log \frac{\sum_w \hat{P}(b,w)}{\sum_w Q(c_t) \prod_{j=1}^n Q(c^j_{t+1}|c^j_t, w) Q(w)} \\
&\leq \inf_{Q \in \mathcal{M}^m_{CII}} \sum_w \sum_b \hat{P}(b,w) \log \frac{\hat{P}(b,w)}{Q(c_t) \prod_{j=1}^n Q(c^j_{t+1}|c^j_t, w) Q(w)} \\
&= \sum_w \sum_b \hat{P}(b,w) \log \frac{\hat{P}(b,w)}{\hat{P}(c_t) \prod_{j=1}^n \hat{P}(c^j_{t+1}|c^j_t, w) \hat{P}(w)} \\
&= \Phi_T.
\end{aligned}$$

By increasing the size of the state space $\mathcal{W}_C$ further, we increase the size of the split model and therefore the value of Φ^m_{CII} can only decrease for a growing m. We discuss the relationship between the split models and how to get from one dimension to the next without increasing Φ^m_{CII} in Section 6.1.

Hence the relation

$$\Phi_{CII} \leq \Phi_T$$

holds.

It follows that if $\Phi_T = 0$ then $\Phi_{CII} = 0$. The log-sum inequality used above is zero if and only if the factors

$$\frac{\hat{P}(b,w)}{\hat{P}(c_t) \prod_{j=1}^n \hat{P}(c^j_{t+1}|c^j_t, w) \hat{P}(w)}$$

are constant. This can be seen in the proof of Jensen's inequality in Theorem 2.6.2 in [Cover06], which is a more general version of the log-sum inequality. Since numerator and denominator are both probability distributions this means they have to be equal and therefore $\Phi_T = 0$. This concludes the proof that both measures are only equal if they are both 0. □

Therefore, by assuming that there exists a common exterior influence, we are able to show that Φ_{CII} is bounded from above by the true value that measures all the intrinsic cross-influences while it is bounded from below by Φ_{CIS}

$$\Phi_T \geq \Phi_{CII} \geq \Phi_{CIS}.$$

3.5.1 Summary of the Relationships among the Measures

Here we summarize the relations among the different Integrated Information measures and their split systems that we discussed in more detail in the previous section.

We start with the split systems. Assuming that Conjecture 1 holds the different split systems are related in the following way:

$$\begin{aligned} \mathcal{M}_I &\subsetneq \mathcal{M}_G \\ \mathcal{M}_{SI} &\subsetneq \mathcal{M}_G \\ \mathcal{M}_I &\subsetneq \mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS} \\ \mathcal{M}_{SI} &\subsetneq \mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS} \end{aligned}$$

A sketch of the inclusion properties among the models is displayed in Figure 3.17 on the left.

This sketch allows us to directly see which postulates the corresponding Integrated Information measures satisfy. Every set that lies inside $\mathcal{M}_{CIS}$ satisfies Postulate 1 and every set that completely contains $\mathcal{M}_I$ fulfills Postulate 2. If Conjecture 1 does not hold then $\mathcal{M}_{CII} = \mathcal{M}_{CIS}$.

Here we did not include $\mathcal{M}_T$ since the elements in $\mathcal{M}_T$ are probability distributions on the larger state space $\mathcal{B} \times \mathcal{W}_C$ instead of $\mathcal{B}$. Therefore the other split models do not directly include or are included in the split system of the ground truth Integrated Information. We would be able to embed $\mathcal{M}_{SI}$ in $\mathcal{M}_T$ by adding a variable W to the elements in $\mathcal{M}_{SI}$ that is independent of B. This is not possible for $\mathcal{M}_G$ or $\mathcal{M}_I$ since the elements in $\mathcal{M}_G$ do not satisfy Postulate 1 and $\mathcal{M}_I$ includes an undirected edge between the $C_{t+1}^j s$ that is not included in $\mathcal{M}_T$.

Nonetheless, as shown in the previous section, we are able to relate Φ_T to the measures Φ_{CII} and Φ_{CIS}, as depicted in the overview on the right of Figure 3.17. There we see that Φ_{SI} and Φ_I are an upper bound for the measures Φ_G, Φ_{CIS} and Φ_{CII}. There is no fixed relation between the measure Φ_G and Φ_{CIS}, Φ_T or Φ_{CII}. Additionally, in Section 4.5.2 Figure 4.10 we see an example where $\Phi_T \geq \Phi_{SI}$.

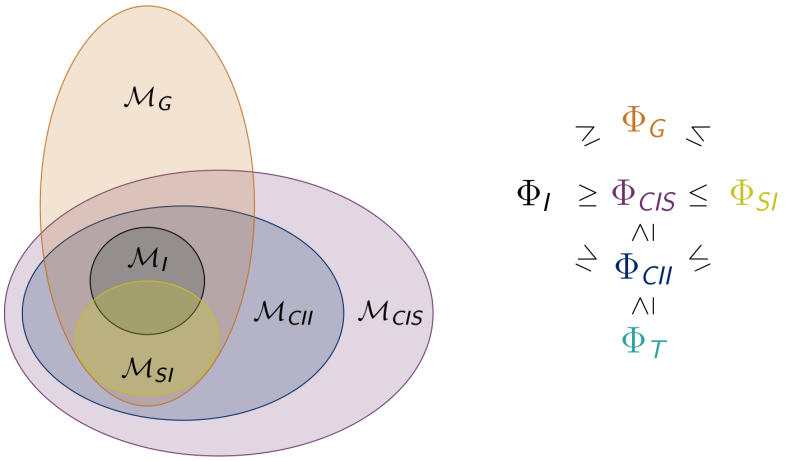


Figure 3.17: Sketch of the relationship between the split systems on the left and the relations between the different Integrated Information measures on the right.

We are be able to observe cases in which Φ_G is greater or smaller than these measures in the next example. There we relate all the different Integrated Information measures, introduced in the previous sections, to each other.

Example 3. In this example we look at a Boltzmann machine inspired system with two binary variables and a binary exterior influence. Note that we calculate the stationary distribution according to V where the exterior influence is included. Therefore this converges to a distribution on $\mathcal{B} \times \mathcal{W}_C$ and afterwards we marginalize to the known variables in order to get $\hat{P}$ for all the measures, except for Φ_T.

$$V = \begin{pmatrix} -0.5 & 0.01 & -1 \\ 0.01 & 0 & 0.3333 \\ 0 & 0 & 0 \end{pmatrix}$$

Figure 3.18 depicts the results for the Integrated Information measures Φ_{SI}, Φ_G, Φ^2_{CII}, Φ^4_{CII}, Φ^8_{CII}, Φ^{16}_{CII}, Φ_T, Φ_{CIS} and Φ_I in case of the inverse temperature β between 0 and 10.

There we observe that Φ_I intersects Φ_{SI} around $\beta = 1$ and Φ_T around $\beta = 5.25$. Therefore, we see that these two measures Φ_{SI} and Φ_T do not satisfy Postulate 2. However, recall that this postulate was justified by considering that there might be an exterior influence on the $C^j_{t+1}s$ that would lead to an undirected edge between them. The measure Φ_T, however, explicitly includes an exterior influence, hence we can argue that Postulate 2 does not apply in this case.

Additionally, the only measure that does not satisfy Postulate 1, Φ_G, intersects Φ_T, Φ_{CIS} and Φ_{CII} around $\beta = 0.5$, $\beta = 4$ and $\beta = 6$.

The results for Φ^m_{CII} are mostly between the curves of Φ_{CIS} and Φ_T, which show an overall similar behavior. Starting roughly at $\beta = 6$ we can observe that the Φ^m_{CII} increase for $m = 16$ even though we have already found better minima in the case of $m = 2, 4$ or 8. Therefore, this leads to the natural question whether it is possible to use the results for lower m in order to guarantee that we will get to a better minimum when we increase the size of the latent space. We explore different approaches to accomplish this in Section 6.1.

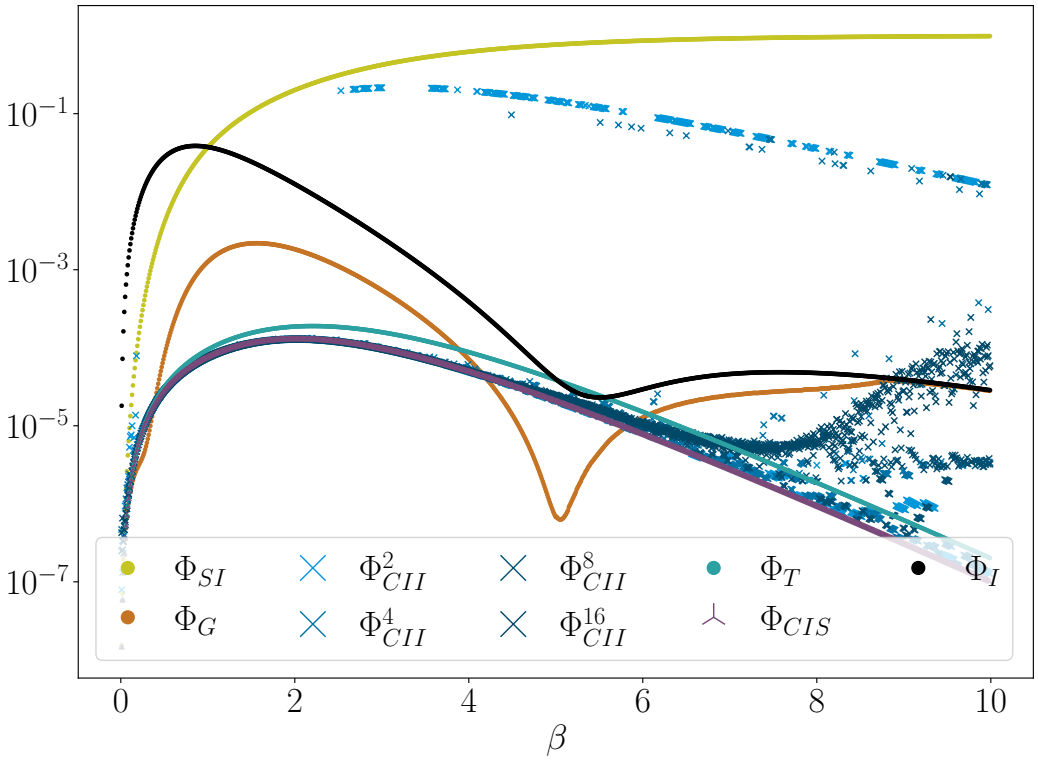


Figure 3.18: The results for the different Integrated Information measures in the case of n=2 and a binary exterior influence.

3.6 Summary and Discussion of Chapter 3

In this chapter we introduce and analyze different measures for Integrated Information. These measures aim at quantifying the causal cross-connections in a system by calculating the KL-divergence between the full distribution and a split model, which is a set of distributions without certain connections.

In [Oizumi16b] and [Amari18] the authors propose two postulates that define properties a valid Integrated Information measure should satisfy. The first postulate identifies conditional independence statements that ensure that the causal cross-connections are removed in the split model. Only one of the discussed measures, namely geometric Integrated Information Φ_G, does not satisfy this condition.

The second postulate requires the Integrated Information measure to be bounded from above by the mutual information Φ_I. The reasoning behind this postulate is that the mutual information removes all the connections between the different points in time and should therefore be an upper bound. However, this no longer holds when there is no exterior influence on our system, as we discuss in more detail after Postulate 2.

One measure that satisfies both postulates is Φ_{CIS}, which is defined via conditional independence statements and has no graphical representation. Hence, the causal nature of the measured information flow is difficult to analyze because of the strong connection between conditional independence statements and graphs in Pearl's causality theory. Additionally, for discrete variables Φ_{CIS} does not have a closed form solution and has to be calculated numerically.

This makes it more difficult to decide whether Φ_{CIS} coincides with our causal information integration measure, called Φ_{CII}. The latter also satisfies both postulates and has additionally a graphical and intuitive interpretation. It is defined by explicitly including an unknown exterior influence on the $C_{t+1}^j s$. Although Φ_{CII} also has no analytical solution we are able to use the information geometric *em*-algorithm to calculate values in case of a fixed-sized latent space. The *em*-algorithm is guaranteed to converge towards a minimum but this might be local. Additionally, by letting the algorithm run multiple times we are able to gain a notion of how the local minima in $\mathcal{E}$ are related to each other as demonstrated in Figure 3.14.

Numerically calculated examples suggest that $\mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS}$. However, even if these two measures coincide we would argue that it might be more reasonable to use Φ_{CII}^m instead. In Example 2 with three binary variables in each point in time even $m = 4096$ is not large enough in order to approximate every element in $\mathcal{M}_{CIS}$. This begs the question whether it is reasonable to use such a huge exterior influence on only three binary variables. Keep in mind that we are still interested in the interior information flow among the C_t and C_{t+1}s. Hence, we would argue that in a setting with an unknown exterior influence we should choose a feasible size m and use Φ_{CII}^m.

These measure should be used in the setting in which there exists an unknown common exterior influence but if we are able to know all the influences on our system, then we are able to calculate Φ_T. This ground truth Integrated Information quantifies truly only the causal cross-connections among C_t^j and $C_{t+1}^{j'}$ for $j \neq j'$. It is defined by including the exterior influence as another known variable in our system. We prove that this true value is an upper bound for Φ_{CII}.

In the following two chapters, Chapter 4 and Chapter 5, we look at small artificial agents that learn to navigate in their environment. Since these agents are purely simulated, we are able to control and know every influence on any part of the controller. This allows us to calculate the actual Integrated Information value, given by Φ_T.

4 The Information Flow in an Acting Agent

In the previous chapter we discuss various Integrated Information measures that solely consider the controller of an agent. Here we introduce a framework in which we can apply the Integrated Information measure Φ_T, the ground truth Integrated Information, to simulated embodied agents. This allows us to compare and understand the Integrated Information value in connection with other information flows inside the agent and between the agent's body and its environment. The results in this chapter were partly published in [Langer21b] and it is structured as follows.

We first introduce the motivation that led to this chapter and highlight our two main results in Section 4.1. One important concept in this context is Morphological Computation, which we introduce in more detail in Section 4.1.1. Then we present the setting of our experiments, in Section 4.2. The goal of these experiments is to analyze the information flow in simple simulated agents in a racetrack. In order to optimize the behavior of the agents we use a technique called "Planning as Inference", explained in Section 4.3. Additionally, we define two subclasses of agents in Section 4.4 before we examine the resulting agents thoroughly with 12 different information theoretic measures, defined in Section 4.5. Afterwards we discuss the results in Section 4.6.

4.1 Introduction

An embodied agent acting in its environment can reach its goals by using solely its brain, its bodies interaction with the world, or a combination of those. Here we want to analyze how embodiment influences the brain and the behavior of an agent. To illustrate the main idea of this chapter consider the following situation.

Imagine a sailor at sea on a boat that lacks navigational equipment. This sailor needs to rely on the information given by his environment, including the sun or the visible stars, in order to determine in which direction he has to steer. This task therefore requires the sailor to combine and process the information from different sources in his brain. However, at the same time a bird equipped with magneto-reception, meaning one that is able to use the magnetic field of the earth to perceive its direction, may rely on this sense and does not need to perform difficult calculations in its brain. Here the body of the bird interacts with the environment and facilitates a better perception. The complexity of the task is met by the morphology of the body of this animal. This leads us to the first of the main statements that we will support with experimental evidence.

1. The more an agent can rely on the interaction of its body with the environment to solve a task, the less Integrated Information in the brain is required.

We use the measure Φ_T as Integrated Information measure to assess the process in the controller, as discussed in the previous Chapter. An introduction to the Integrated Information Theory of consciousness is given in Section 3.1. The interaction of an agent's body with its environment that allows for fewer calculations in the brain is called Morphological Computation. We introduce this concept in more detail in the following section.

Coming back to our example, a modern boat with a highly developed navigation system would also alleviate the mental load of its crew. The person in charge now only needs to understand how to use and interpret the navigation devices. The complexity of the task shifts from the brain and background knowledge of a person towards the construction of the navigation system. Certainly, the brain of this sailor does not necessarily integrate less information in this scenario. He might pass the time by doing complicated puzzles, which do not have anything to do with his navigation. Therefore, we need to differentiate between

behavioral relevant and irrelevant information integration. Hence, our second result is the following.

2. The importance of Integrated Information in the controller for the behavior of an embodied agent depends additionally on the information flowing to and from the controller. Therefore it is not sufficient to only calculate an Integrated Information measure.

This statement is supported by the observation that the antagonistic relationship between Integrated Information and Morphological Computation exists even in cases in which the controller has no influence on the behavior of the agent at all. Hence it is necessary to further analyze the information flow in order to fully understand the impact the controller has on the behavior of the agent. We emphasize this point by defining a new measure, the effective information integration.

Before we introduce our experiments in detail, we first discuss different aspects of the concept of Morphological Computation.

4.1.1 Morphological Computation

The notion that the body of an agent is crucial for its movement and behavior has been long established in the study of motor control. In his influential work, published in english in 1967 [Bernstein67], the neurophysiologist Bernstein addresses the difficulties resulting from the many degrees of freedom within a human body. In [Bernstein67], in the chapter "Conclusions towards the study of motor co-ordination", he makes the following observation:

> "All these many sources of indeterminacy lead to the same end result; which is that the *motor effect of a central impulse cannot be decided at the centre* but is decided entirely at the periphery: at the last spinal and myoneural synapse, at the muscle, in the mechanical and anatomical change of forces in the limb being moved, etc. "

Thereby, he emphasizes that the brain is not fully determining the movement, but that the interaction of the body with its environment governs the outcome.

This concept was demonstrated in 1986 by Brooks who describes in [Brooks86] a layered control system that decomposes into parallel processing and can run a robot with a complex morphology in real time. In his following works [Brooks91b; Brooks91a] he advocates for using an embodied robot situated in a real environment and acting in real-time in order to built intelligent systems.

One important aspect resulting from the interaction of an agent's body with its environment is called Morphological Computation and was defined in [Pfeifer06a]. This describes the phenomenon that occurs when the interaction of the agent's body with its environment performs functions that otherwise could have been done by the controller, the brain of the agent.

The exact definition of the term "Morphological Computation" differs between authors and changed over time, as explained in [Ghazi-Zahedi19] in Section 1.2.1. The main point of disagreement stems from the use of the word "computation". Some authors understand this term as actual computing in the sense of a Turing machine. In [Füchslin13] the authors refer to a workshop at the first International Conference on Morphological Computing in Venice in 2007 in which Morphological Computation was defined in such a manner. They describe 3 properties that a process should satisfy in order to be called Morphological Computation, namely a process that "(a) serves for a computational purpose, (b) has clearly assignable input and output states and (c) is programmable, where 'programmable' is understood in

the broad sense that a programmer can vary the behaviour of the system by varying a set of parameters.", [Füchslin13]. This strict definition excludes many examples that are widely considered to demonstrate Morphological Computation.

A famous example that falls into this category is the passive dynamic walker. This is a purely mechanical robot, published in 1990 in [McGeer90a; McGeer90b], that can move down a decline without any control architecture. It is an extreme case exemplifying the cheap design principle in embodied artificial intelligence, as described in [Pfeifer01; Montúfar15]. Inspired by the passive dynamic walker there exist various examples of robots that exploit their materials and dynamics in order to achieve a rich behavior with only a very limited controller. Three of these examples are given in [Pfeifer09], which are a walking and hopping robot "stumpy", a robot walking on four legs equipped with springs named "puppy" and "crazy bird", a quadrupedal robot. Müller and Hoffmann argue in [Müller17] that these types of examples do not classify as Morphological Computation because the body is not actually computing. In regards to the passive dynamic walker they ask the question: "If this is computing, what is not?".

A more detailed discussion on this issue can be found in [Ghazi-Zahedi19]. In order to resolve this conflict the author suggests to refer to Morphological Computation as "Morphological Intelligence" and characterizes it in Definition 1.1. as follows

> "Morphological Intelligence is the reduction of computational cost for the brain (or controller) resulting from the exploitation of the morphology and its interaction with the environment."

Note that this definition does not reference "computing" in connection with the body at all. Instead of using this unifying definition Müller and Hoffmann suggest to prevent further confusion by distinguishing between three different categories. These are

(1) Morphology facilitating control,

(2) Morphology facilitating perception and

(3) proper Morphological Computation.

We now discuss these three categories in more detail starting with proper Morphological Computation. An instance of this category can be found in reservoir computing. This term describes a framework in which a complex system is used as a computational medium where only a readout layer or function is learned. Different types of reservoirs include, for example, recurrent neural networks or the bodies of soft robots, as discussed in [Baldini22]. Hence, the field of soft robotics where the tissue of the robot's bodies allows for a simple control architecture contains examples of proper Morphological Computation. These are described, for example, in [Nakajima13a; Nakajima13b; Nakajima14]. However, a robot with a more complex morphology might also bring disadvantages, as pointed out in [Hoffmann17; Ghazi-Zahedi17b].

The second category "Morphology facilitating perception" is called "pre-processing" in [Ghazi-Zahedi19]. The properties of the sensors of an agent directly influence how it perceives its environment and therefore how well it can interact with it. This can heavily impact the complexity of the task an agent is facing. One example of well-designed sensors are the compound eyes of flies, which have been analyzed and used for building an obstacle avoiding robot in [Franceschini92]. In our experiments, in the following sections, we manipulate the sensors of the agents in order to influence their ability to interact with their environment.

Similarly, agents can make use of the properties of their body to simplify the required action commands, which is called "post-processing" in [Ghazi-Zahedi19]. The complexity of the design of muscles, for example, directly impacts the amount of necessary control,

discussed in [Häufle10; Häufle14; Ghazi-Zahedi16]. Another example of post-processing is the flexible spine of a cat, mentioned in the introduction in Section 1.1.

The last category "Morphology facilitating control", or equivalently "mechanical control" in [Hoffmann17], includes the passive dynamic walker and all those systems that utilize the interaction of their bodies with their environment without having a computational level. Additionally, the authors of [Müller17] include the example of gecko feet in this context. The morphology of gecko feet facilitates the use of van der Waals adhesion and therefore allows them to stick to smooth surfaces, described in [Autumn02]. Note that this case is also excluded in the "Morphological Intelligence" definition. In [Ghazi-Zahedi19] the author terms these body-environment interactions as "Behavior-Enabling Physical Processes". In these cases a controller could not compensate if the physical process would not exist.

These variations in terminology and conceptional perspective encourage the author of [Miłkowski18] to a harsh criticism of the concept of Morphological Computation. He concludes that "the notion may be confusing in the worst cases. At best, however, it is nothing but physical computation." Nonetheless, this is still an active field of research and we do expect that there is additional value in explicitly highlighting the impact of the morphology of an agent on the necessary complexity of its control architecture. We understand Morphological Computation as the interaction of an agent's body with its environment that eases the computational burden of the controller.

Additionally, the question of how to measure Morphological Computation arises. There exist various measures for it. In [Klyubin05] the authors define an information theoretic measure called "empowerment" that quantifies how large the influence of the agent is on its next sensory inputs. Furthermore, the authors of [Ghazi-Zahedi13] and [Ghazi-Zahedi17a] define different measures considering the information flow among the sensors, the actuators and the world. These measures are compared thoroughly in [Ghazi-Zahedi19] and we introduce three of them in Section 4.5.3.

4.2 The Setting of the Experiments

We examine the information flow of agents that act in a simple environment. The agents are idealized models of a two-wheeled robots depicted in Figure 4.1 on the left. Each wheel can spin either fast or slow, hence the agents have four different possible movements. If both wheels spin fast, then the agent moves 0.6 units of length forward and if they both spin slow, then the agent moves 0.2. In case of one fast and one slow wheel the agent makes a turn of approximately 10° with a speed of 0.4 units of length per step. Note that the agents are not able to stop. The code of the movement and a video of 5 agents performing random movements can be found in [Langer21a].

The agent's body consists of a blue circle and a blue line marking its back. Additionally, the agents have two black lines that are binary sensors, meaning that they only detect whether they touch an obstacle or not. They do not report the exact distance to a wall. If a sensor touches a wall it turns green and if the body of the agent touches a wall it turns red. In the middle of Figure 4.1 multiple agents are depicted in their environment, a racetrack.

In this racetrack the agents are faced with the challenge to move without letting their bodies touch the walls. An agent that fails to avoid the walls turns red and dies. Hence the goal for the agents is to stay alive. The design and first implementation of the agents and the racetrack is due to Virgo, [Virgo19].

Note that although we depicted more than one agent in the environment these agents have no impact on each other.

Additionally, we want to manipulate the morphology of the agents and we do that by varying the length of the sensors. Thereby we can directly influence the agent's perception

of its environment and the potential of interacting with it. Hence, changing the reach of the sensors has an immediate influence on the Morphological Computation. In Section 4.1.1 this concept was discussed as "Morphology facilitating perception". We vary the reach of the sensors from 0.5 to 2.75. Four different sensor length are depicted in Figure 4.1 on the right.

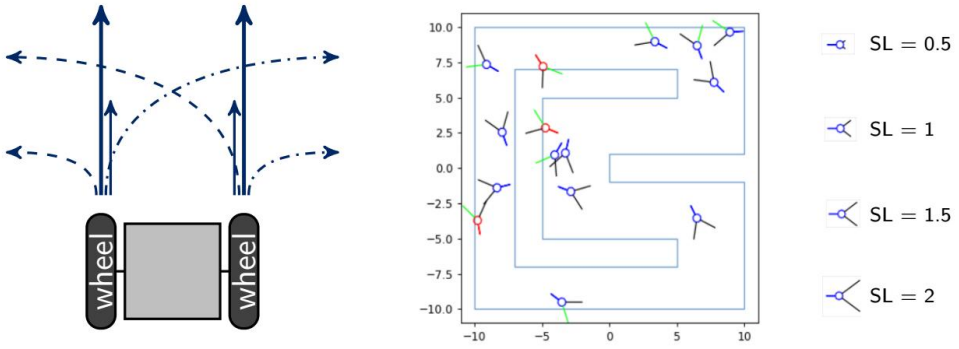


Figure 4.1: (left) A sketch of a two-wheeled robot and its four different types of movement. (middle) The racetrack the agents have to survive in and the different sensor lengths, named SL, on the right. [Langer21b]

This manipulation can also alter the strategies of the agents. The authors of [Maris96] analyze the behavior of real robots, called Didabots, that are similar to our simulated ones but with a fixed control algorithm. In [Pfeifer98] Box 2, the authors note that a change in the sensor position can modify the behavior of the robots completely even though the commands from the controller stay the same.

We calculate the strategies that the agents use by applying the concept of planning as inference as discussed in Section 4.3. Utilizing this method allows us to directly determine the optimal behaviors without having to train any agents first.

Before we discuss this further we present the probabilistic architecture of the agents and their environment in the next section.

4.2.1 The Agents

The behavior of an embodied agent acting in its environment can be modeled by the sensorimotor loop, depicted in Figure 4.2. There we see the different information flows among the world, the actuators, the sensors and the controller. The agent perceives information about the world through its sensors, which can send them as a direct stimuli-response to the actuators. Additionally, the sensors send the information to the controller that can process this information leading to commands being send to the actuators. The states of the actuators then lead to actions of the agent and thereby influence the environment, which in turn affects the sensors. This is also called action-perception cycle, discussed in more detail in Section 1.2 and in, for example, [Klyubin04; Ay14; Ay15b].

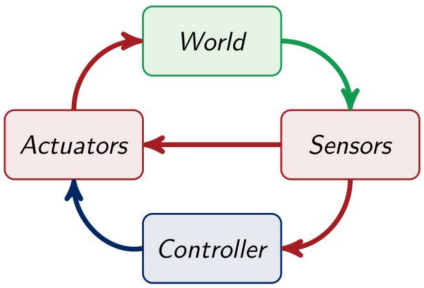


Figure 4.2: The sensorimotor loop.

Now we focus on one time step from t to $t+1$. Unfolding the relations among the world, controller, sensors and actuators in time leads to the graph in Figure 4.3. The agents in our setting have two binary sensory S_t^k, $k \in \{1,2\}$, two binary controller C_t^j, $j \in \{1,2\}$ and two binary actuator nodes A_t^i, $i \in \{1,2\}$. The sensor and controller nodes send their information to the actuator and controller nodes in the next point in time. The sensors are only influenced by the world W and the world is only affected by the actuators and the last world state as depicted in Figure 4.3.

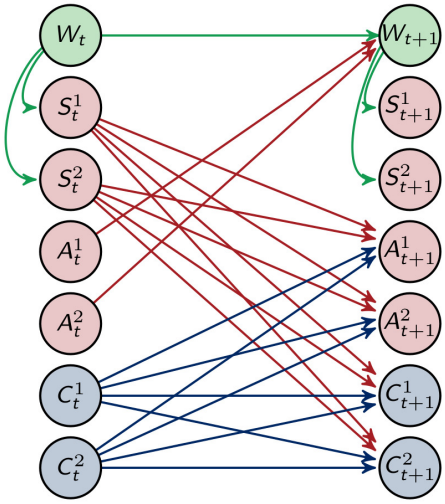


Figure 4.3: The architecture of the agents.

The behavior of the agents is governed by a probabilistic law that can be modeled as the following discrete, multivariate, time-homogeneous Markov-Process

$$(W_t, X_t)_{t\in\mathbb{N}} = (W_t, S_t, A_t, C_t)_{t\in\mathbb{N}}$$

with the state space $\mathcal{W} \times \mathcal{X} = \mathcal{W} \times \mathcal{S} \times \mathcal{A} \times \mathcal{C}$.

The distribution describing the Markov-Process can be written as

$$P(w_0, x_0, \ldots, w_{t+1}, x_{t+1}) = P(w_0, x_0) \prod_{\ell=0}^{t} P(w_{\ell+1}, x_{\ell+1} | w_\ell, x_\ell),$$

for all $(w_\ell, x_\ell) \in \mathcal{W} \times \mathcal{X}$ with $\ell \in \{0, \ldots, t\}$. There the transition probability from ℓ to $\ell+1$ is factorized as follows

$$P(w_{\ell+1}, x_{\ell+1} | w_\ell, x_\ell) = P(w_{\ell+1} | w_\ell, a_\ell) \prod_k P(s_{\ell+1}^k | w_{\ell+1}) \prod_i P(a_{\ell+1}^i | s_\ell, c_\ell) \prod_j P(c_{\ell+1}^j | s_\ell, c_\ell),$$

for all $(w_\ell, x_\ell) \in \mathcal{W} \times \mathcal{X}$.

The directed acyclic graph corresponding to this factorization is depicted in Figure 4.4 on the top. To simplify we only draw one node for each S, A and C in Figure 4.4 and the following depictions. An introduction to graphs and their associated graphical models is given in Section 2.4 and more details can be found in [Lauritzen96].

Throughout this chapter we assume that the distributions on $\mathcal{X}$, that define the agents, are strictly positive unless stated otherwise. This ensures that the conditional probability distributions in the factorization above are well defined.

In the next section we take a closer look at the role of the environment.

4.2.2 The Environment

The previous section introduces the Markov-Process $(W_t, X_t)_{t\in\mathbb{N}}$ that describes the agent and its environment. The agent itself has access to its actuator and sensory states, but it cannot observe the world W directly. It has only access to the world through the sensors and therefore we replace the distributions $P(W_{t+1}|W_t, A_t)$ and $P(S_{t+1}^k|W_{t+1})$, for $k \in \{1,2\}$, using exclusively information intrinsically known to the agent. In order to do that we look closer at one step in time

$$P(w_t, x_t, w_{t+1} x_{t+1}) = P(x_t, w_t) P(x_{t+1}, w_{t+1} | x_t, w_t),$$

for all $(x_t, w_t, x_{t+1}, w_{t+1}) \in \mathcal{X} \times \mathcal{W} \times \mathcal{X} \times \mathcal{W}$.

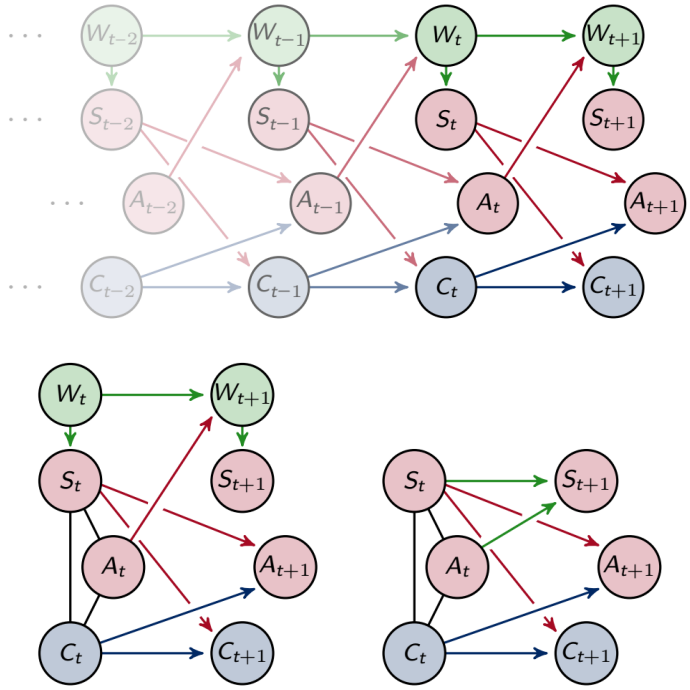


Figure 4.4: The graphical representation of the Markov-Process $(W_t, X_t)_{t\in\mathcal{N}}$ on the top, the graphical representation of one timestep on the bottom left and the marginalized graph on the bottom right.

Here we additionally assume that the environment only influences the sensors even in the graph marginalized to one timestep as depicted in Figure 4.4 on the bottom left. In the next proposition we use this assumption and sum over $w_t, w_{t+1} \in \mathcal{W}$ leading to a Markov-Process that only depends on the variables known to the agent.

Proposition 9. *Let P be a distribution that factorizes according to the graph on the bottom left of Figure 4.4, that is*

$$P(w_t, x_t, w_{t+1}, x_{t+1}) = P(w_t)P(x_t|w_t)P(w_{t+1}|w_t, a_t)\prod_k P(s^k_{t+1}|w_{t+1}) \cdot \prod_i P(a^i_{t+1}|s_t, c_t)\prod_j P(c^j_{t+1}|s_t, c_t),$$

for all $(x_t, w_t, x_{t+1}, w_{t+1}) \in \mathcal{X}\times\mathcal{W}\times\mathcal{X}\times\mathcal{W}$. Marginalizing this distribution over the world states $(w_t, w_{t+1}) \in \mathcal{W}\times\mathcal{W}$ leads to the following Markov-Process

$$P(x_t, x_{t+1}) = P(s_t, a_t, c_t)\prod_i P(a^i_{t+1}|s_t, c_t)\prod_j P(c^j_{t+1}|s_t, c_t)P(s_{t+1}|s_t, a_t).$$

Proof of Proposition 9. Marginalizing over the influences from the world leads to

$$\begin{aligned} &P(x_t, x_{t+1}) \\ &= \sum_{w_t, w_{t+1}} P(w_t)P(x_t|w_t)P(w_{t+1}|w_t, a_t)\prod_k P(s^k_{t+1}|w_{t+1})\prod_i P(a^i_{t+1}|s_t, c_t)\prod_j P(c^j_{t+1}|s_t, c_t) \\ &= \prod_i P(a^i_{t+1}|s_t, c_t)\prod_j P(c^j_{t+1}|s_t, c_t)\sum_{w_t, w_{t+1}} P(w_t)P(x_t|w_t)P(w_{t+1}|w_t, a_t)\prod_k P(s^k_{t+1}|w_{t+1}), \end{aligned}$$

for all $(x_t, x_{t+1}) \in \mathcal{X} \times \mathcal{X}$. The sum describes

$$P(s_t, a_t, c_t, s_{t+1}) = \sum_{w_t, w_{t+1}} P(w_t)P(x_t|w_t)P(w_{t+1}|w_t, a_t) \prod_k P(s^k_{t+1}|w_{t+1}).$$

Now we take a closer look at $P(S_{t+1}|S_t, A_t, C_t)$ and show that S_{t+1} is independent of C_t given (S_t, A_t). For that we need to describe the distribution $P(X_t, W_t)$ in more detail. The graph corresponding to this distribution is a chain graph and therefore we are able to use the parametrization described in Section 2.4.3. There exist non-negative functions f_1, f_2, such that

$$P(s_t, a_t, c_t) = f_2(s_t, a_t, c_t) \sum_w f_1(s_t, w_t),\ \forall (x_t, w_t) \in \mathcal{X} \times \mathcal{W}.$$

Using this definition results in

$$\begin{aligned} P(s_{t+1}|s_t, a_t, c_t) &= \frac{P(s_{t+1}, s_t, a_t, c_t)}{P(s_t, a_t, c_t)} \\ &= \frac{f_2(s_t, a_t, c_t) \left(\sum_{w_t} f_1(s_t, w_t) \sum_{w_{t+1}} P(w_{t+1}|w_t, a_t) \prod_k P(s^k_{t+1}|w_{t+1}) \right)}{f_2(s_t, a_t, c_t) \sum_{w_t} f_1(s_t, w_t)} \\ &= \frac{\sum_{w_t} f_1(s_t, w_t) \sum_{w_{t+1}} P(w_{t+1}|w_t, a_t) \prod_k P(s^k_{t+1}|w_{t+1})}{\sum_{w_t} f_1(s_t, w_t)}, \end{aligned}$$

for all $(s_{t+1}, x_t) \in \mathcal{S} \times \mathcal{X}$. Similarly, the following holds

$$\begin{aligned} P(s_{t+1}|s_t, a_t) &= \frac{\left(\sum_{c_t} f_2(s_t, a_t, c_t) \right) \left(\sum_{w_t} f_1(s_t, w_t) \sum_{w_{t+1}} P(w_{t+1}|w_t, a_t) \prod_k P(s^k_{t+1}|w_{t+1}) \right)}{\left(\sum_{c_t} f_2(s_t, a_t, c_t) \right) \sum_{w_t} f_1(s_t, w_t)} \\ &= \frac{\sum_{w_t} f_1(s_t, w_t) \sum_{w_{t+1}} P(w_{t+1}|w_t, a_t) \prod_k P(s^k_{t+1}|w_{t+1})}{\sum_{w_t} f_1(s_t, w_t)}, \end{aligned}$$

for all $(s_{t+1}, s_t, a_t) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$. Therefore $P(S_{t+1}|S_t, A_t) = P(S_{t+1}|S_t, A_t, C_t)$, hence $S_{t+1} \perp\!\!\!\perp C_t|(S_t, A_t)$ and the factorization of P can be written as

$$P(x_t, x_{t+1}) = P(x_t) \prod_i P(a^i_{t+1}|s_t, c_t) \prod_j P(c^j_{t+1}|s_t, c_t) P(s_{t+1}|s_t, a_t),\ \forall (x_t, x_{t+1}) \in \mathcal{X} \times \mathcal{X}.$$

□

The new process defined in Proposition 9 describes the behavior of the environment only with information known by the agent and is shown in Figure 4.4 on the bottom right. A similar distribution is also used in Section 3.3.1. of [Ghazi-Zahedi13] and in [Ghazi-Zahedi19]. There it is derived by taking $P(S_{t+1}|S_t)$ as the intrinsically available information of $P(W_{t+1}|W_t)$. In that case the distribution $P(S_{t+1}|S_t, A_t)$ is called "internal world model". This distribution plays an important role in Chapter 5 and we discuss it in more detail in Section 4.2.1 where we name it "empirical world model".

Here we sample the distribution $\tilde{P}(S_{t+1}, S_t, A_t)$ for every sensor length. To that end we store 20 000 000 sensor and motor values for agents that start in a random place in the arena and perform arbitrary movements. We denote all the sampled and therefore fixed distributions by $\tilde{P}$.

Since we are now able to define a set of distributions that describe the interaction between the agent and the world according to the sensorimotor loop, we present the method to find the optimal behavior in the next section.

4.3 Optimizing the Behavior via Planning as Inference

In this section we discuss how to determine the behavior of the agents. We calculate the optimal behavior by using the concept of "Planning as Inference". This is a theory for planning under uncertainty and it was originally proposed by Attias in [Attias03] and further developed by Toussaint and collegues in [Toussaint06; Toussaint08; Toussaint09]. There the conditional distributions that define the actions of the agents are considered to be latent variables that have to be optimized.

The optimization is done by using the EM-algorithm, which is equivalent to the information theoretic *em*-algorithm in this case. We describe the *em*-algorithm in detail in Section 2.5.1 and address its usage in this context in the next section. The *em*-algorithm is well known and was proposed in 1984 in [Csiszár84], further discussed in [Amari95; Amari92]. The resulting distribution maximizes the likelihood of achieving the predefined goal but might lead to a local optimum, as discussed in the context of causal information integration in Section 3.4.1. Normally this is a disadvantage but in our setting it allows us to analyze various strategies by using different initial distributions.

The goal of an agent in our experiment is to maximize the probability of being alive after the next two movements. We need to include two steps instead of one in order make sure that the connection between C_t and C_{t+1} has an impact on the outcome. This can be seen in Figure 4.5.

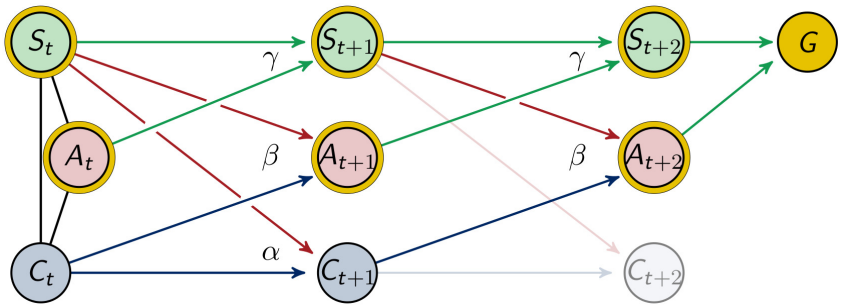


Figure 4.5: Graphical representation of two timesteps.

We denote the goal variable by G with the state space $\mathcal{G} = \{0, 1\}$ where the value $P(g = 1) := P(g_1)$ refers to the probability of the agent to be alive and $P(g = 0)$ is the probability of the agent to be dead. Since the agent moves twice, this distribution depends on the states of the last three sensor and actuator states, which are marked by a golden circle in Figure 4.5

$$\tilde{P}(G|S_{t+2}, S_{t+1}, S_t, A_{t+2}, A_{t+1}, A_t).$$

Note that this is a fixed, sampled distribution hence it is denoted by a tilde. We sample this distribution for every sensor length as described in the previous section in the context of $\tilde{P}(S_{t+1}, S_t, A_t)$.

The architecture of the agents is discussed in the previous sections. There we outlined how we sample the distribution $\gamma := \tilde{P}(S_{t+1}|S_t, A_t)$ that describes the influence the agent has on itself through the world. The distributions that define the behavior of the agents are

$$\beta := P(A_{t+1}|S_t, C_t) \qquad \text{and} \qquad \alpha := P(C_{t+1}|S_t, C_t).$$

Hence, we treat (A_{t+1}, C_{t+1}) as latent variables and optimize their distributions with respect to the goal of maximizing $P(g_1)$. We denote these distributions by α, β and γ in order to emphasize that the process is time-homogeneous, meaning that

$$\begin{aligned} P(A_{t+1}|S_t, C_t) &= P(A_{t+2}|S_{t+1}, C_{t+1}) \\ P(S_{t+1}|S_t, A_t) &= P(S_{t+2}|S_{t+1}, A_{t+1}) \\ P(C_{t+1}|S_t, C_t) &= P(C_{t+2}|S_{t+1}, C_{t+1}) \end{aligned}$$

as indicated in Figure 4.5. Note that the above mentioned homogeneity does not imply stationarity.

It remains to define the initial distribution $P(S_t, C_t, A_t) = P(X_t)$. This can be written in the following way

$$P(x_t) = P(c_t|a_t, s_t)P(s_t|a_t)P(a_t),$$

for all $x_t \in \mathcal{X}$.

Now we use the sampled distribution $\tilde{P}(S_{t+1}, S_t, A_t)$ so that we are able to calculate $\tilde{P}(S_t|A_t)$ and set $P(S_t|A_t) = \tilde{P}(S_t|A_t)$. The remaining distributions $P(C_t|A_t, S_t)$ and $P(A_t)$ are also treated as hidden variables and optimized using the *em*-algorithm, as described in the next section. This approach leads to the optimal starting conditions for the agents.

4.3.1 The em-Algorithm in the Context of Planning as Inference

In this section we define the optimization in the context of planning as inference in more detail. This optimization was introduced in general in the previous section and it uses the *em*-algorithm.

Before we define the steps of the *em*-algorithm we first introduce some new notation to simplify the following steps. We define the state space $\mathcal{Z} = \mathcal{X} \times \mathcal{X} \times \mathcal{S} \times \mathcal{A}$, such that

$$z = (s_t, c_t, a_t, s_{t+1}, c_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}) \in \mathcal{Z}.$$

Here C_{t+2} is not included since its value has no influence on the likelihood of the goal as one can see in Figure 4.5 where C_{t+2} is depicted as faded. We refer to all the sampled distributions regarding the environment by

$$\tilde{P}_{s|s,a} = \tilde{P}(s_t|a_t)\tilde{P}(s_{t+1}|s_t, a_t)\tilde{P}(s_{t+2}|s_{t+1}, a_{t+1})\tilde{P}(g|s_{t+2}, s_{t+1}, s_t, a_{t+2}, a_{t+1}, a_t),$$

for all $(z, g) \in \mathcal{Z} \times \mathcal{G}$.

The *em*-algorithm iterates between two sets of distributions in order to find the minimal difference between them, as described in Section 2.5.1. The first of these sets is called the goal manifold, $\mathcal{M}_G$, since every element in this set achieves the goal with probability one:

$$\mathcal{M}_G = \{P \in \mathcal{P}(\mathcal{Z} \times \mathcal{G}) \,|\, P(g_1) = 1\}.$$

The second set consists of all the distributions that factor according to the architecture of the agents, meaning each of these distributions describes a valid behavior of an agent.

Therefore, we call this the agent manifold, $\mathcal{M}_A$:

$$\mathcal{M}_A = \Bigg\{ Q \in \mathcal{P}^\circ(\mathcal{Z} \times \mathcal{G}) \,|\, Q(z,g) = \tilde{P}_{s,g|s,a}\, Q(c_t|a_t,s_t)Q(a_t) \\ \prod_i Q(a^i_{t+1}|s_t,c_t)Q(a^i_{t+2}|s_{t+1},c_{t+1}) \prod_j Q(c^j_{t+1}|s_t,c_t),\ (z,g) \in \mathcal{Z} \times \mathcal{G} \Bigg\}.$$

The elements in the set $\mathcal{M}_A$ are not time-homogeneous since we did not explicitly include that

$$Q(A^i_{t+1}|S_t,C_t) = Q(A^i_{t+2}|S_{t+1},C_{t+1}), \quad i \in \{1,2\}.$$

We define the set in this way because we are able to perform an exact m-projection to $\mathcal{M}_A$, as we show below. In order to assure time-homogeneity we have to make an additional assumption, which leads to an approximate minimization.

Note that $\mathcal{P}(\mathcal{Z})$ is the set of probability distributions with the state space $\mathcal{Z}$ and $\mathcal{P}^\circ(\mathcal{Z})$ consists of all the strictly positive distributions in $\mathcal{P}(\mathcal{Z})$. Hence, the two manifolds, $\mathcal{M}_G$ and $\mathcal{M}_A$, are disjoint. Every distribution in $\mathcal{M}_G$ has per definition values equal to zero and is therefore on the boundary of the probability simplex, whereas the elements in $\mathcal{M}_A$ are positive probability distributions.

The *em*-algorithm works by iteratively performing e-projections to $\mathcal{M}_G$ and m-projections to $\mathcal{M}_A$. This leads to the minimal difference between elements of these two manifolds with respect to the KL-divergence

$$\inf_{P \in \mathcal{M}_G, Q \in \mathcal{M}_A} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q).$$

Therefore this procedure results in the distribution $Q^\star \in \mathcal{M}_A$, which describes a valid agent and is closest to achieving the goal.

In the remainder of this section we define the e- and m-projections for the sets $\mathcal{M}_G$ and $\mathcal{M}_A$ and we start with the e-projection. An e-projection of $Q \in \mathcal{M}_A$ to the linear family $\mathcal{M}_G$ is well known and can be performed in the following way

$$\begin{aligned} \arg\min_{P \in \mathcal{M}_G} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q) &= P^\star \\ P^\star(z,g) &= Q(z,g)\frac{P(g)}{Q(g)}, \end{aligned} \tag{4.1}$$

for all $(z,g) \in \mathcal{Z} \times \mathcal{G}$.

Note that $P(g)$ is the same for every element in $\mathcal{M}_G$ and that $Q \in \mathcal{M}_A$ is strictly positive. Therefore this expression is well defined.

Proof of (4.1). In order to prove that $P^\star$ is the e-projection of Q to $\mathcal{M}_G$ we make use of the log-sum inequality. Let $P \in \mathcal{M}_G$ then

$$\begin{aligned} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q) &= \sum_{z,g} P(z,g)\, \log\left(\frac{P(z,g)}{Q(z,g)}\right) \\ &\geq \sum_g \left(\sum_z P(z,g)\right) \log\left(\frac{\sum_z P(z,g)}{\sum_z Q(z,g)}\right). \end{aligned}$$

Now we use convention that $0 \cdot \log 0 = 0$ and this leads to

$$\begin{aligned} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q) &\geq 1 \cdot \log\left(\frac{1}{Q(g_1)}\right) + 0 \cdot \log\left(\frac{0}{Q(g_0)}\right) \\ &= \sum_z Q(z, g_1)\frac{1}{Q(g_1)} \log\left(\frac{1 \cdot Q(z, g_1)}{Q(g_1)Q(z, g_1)}\right) \\ &\quad + Q(z, g_0)\frac{0}{Q(g_0)} \log\left(\frac{Q(z, g_0) \cdot 0}{Q(g_0)}\right) \\ &= \sum_z P^\star(z, g_1) \log\left(\frac{P^\star(z, g_1)}{Q(z, g_1)}\right) + P^\star(z, g_0) \log\left(\frac{P^\star(z, g_0)}{Q(g_0)}\right) \\ &= D_{\mathcal{Z}\times\mathcal{G}}(P^\star \parallel Q) \end{aligned}$$

□

Next we define the m-projection of $P \in \mathcal{M}_G$ to $\mathcal{M}_A$. Minimizing the KL-divergence with respect to an element $Q \in \mathcal{M}_A$ leads to

$$\begin{aligned} &\arg\min_{Q\in\mathcal{M}_A} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q) = Q^\star \\ &\qquad Q^\star(z, g) = P(c_t|a_t, s_t)\tilde{P}(s, g|s, a)P(a_t)\prod_j P(c^j_{t+1}|s_t, c_t) \\ &\qquad\qquad \prod_i P(a^i_{t+1}|s_t, c_t)P(a^i_{t+2}|s_{t+1}, c_{t+1}),\ \forall (z, g) \in \mathcal{Z} \times \mathcal{G}. \end{aligned} \tag{4.2}$$

We discuss the matter of the time-homogeneity after the proof of this projection.

Proof of (4.2). The KL-divergence between $Q \in \mathcal{M}_A$ and $P \in \mathcal{M}_G$ can be written as

$$\begin{aligned} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q) &= \sum_{z,g} P(z, g) \log\frac{P(z, g)}{Q(z, g)} \\ &= \sum_z P(z, g) \log\frac{P(z, g)}{\tilde{P}(s, g|s, a)} + \sum_{z,g} P(z, g) \log\frac{1}{Q(c_t|a_t, s_t)} \\ &+ \sum_{z,g} P(z, g) \log\frac{1}{Q(a_t)} + \sum_{z,g} P(z, g) \log\frac{1}{\prod_i Q(a^i_{t+1}|s_t, c_t)} \\ &+ \sum_{z,g} P(z, g) \log\frac{1}{\prod_i Q(a^i_{t+2}|s_{t+1}, c_{t+1})} + \sum_{z,g} P(z, g) \log\frac{1}{\prod_j Q(c^j_{t+1}|s_t, c_t)}. \end{aligned}$$

The last five sums are cross-entropies, as defined in Section 2.2. Since the cross-entropy is greater or equal to the entropy, we gain the following inequality.

$$\begin{aligned} D_{\mathcal{Z}\times\mathcal{G}}(P \parallel Q) &\geq \sum_{z,g} P(z, g) \log\frac{P(z, g)}{\tilde{P}(s, g|s, a)} + \sum_{z,g} P(z, g) \log\frac{1}{P(c_t|a_t, s_t)} \\ &+ \sum_{z,g} P(z, g) \log\frac{1}{P(a_t)} + \sum_{z,g} P(z, g) \log\frac{1}{\prod_i P(a^i_{t+1}|s_t, c_t)} \\ &+ \sum_{z,g} P(z, g) \log\frac{1}{\prod_i P(a^i_{t+2}|s_{t+1}, c_{t+1})} + \sum_{z,g} P(z, g) \log\frac{1}{\prod_j P(c^j_{t+1}|s_t, c_t)} \\ &= D(P \parallel Q^\star) \end{aligned}$$

□

Now, we defined the e- and m-projections to the sets $\mathcal{M}_G$ and $\mathcal{M}_A$ respectively and we can apply the em-algorithm. However, as mentioned earlier, it remains to ensure that the resulting distribution is time-homogeneous. This is not guaranteed since in the proof of the m-projection above we use that the cross-entropy is greater or equal to the entropy and therefore we arrive at the following distributions for the projections

$$Q(A^i_{t+1}|S_t, C_t) = P(A^i_{t+1}|S_t, C_t)$$
$$Q(A^i_{t+2}|S_{t+1}, C_{t+1}) = P(A^i_{t+2}|S_{t+1}, C_{t+1}).$$

Since $P \in \mathcal{M}_G$ these distributions are in general not equal. Therefore we choose to approximate the projection by setting

$$Q(A^i_{t+2}|S_{t+1}, C_{t+1}) := Q(A^i_{t+1}|S_t, C_t) = P(A^i_{t+1}|S_t, C_t).$$

By doing so we are replacing a cross-entropy by another cross-entropy. Hence, we are no longer guaranteed that this leads to the minimum, the exact projection. However, this ensures that the resulting distribution is time-homogeneous and in practice we observe a decreasing sequence of KL-divergences. The experimental results, discussed in Sections 4.5 and 4.6 were produced by performing these steps for at least 1000 iterations and until the improvement in the probability of reaching the goal is smaller than 10^{-5}. Hence, although we are not able to guarantee a convergence in this modified case the practical application of this algorithm shows satisfying results.

4.4 The Three Distinct Types of Agents

In order to analyze the information flow of the acting agents thoroughly we define two additional types of agents. The ones previously introduced in Section 4.2.1 are called "fully coupled" and they are factorized by the following distribution

$$P(x_t, x_{t+1}) = P(s_t, a_t, c_t) \prod_i P(a^i_{t+1}|s_t, c_t) \prod_j P(c^j_{t+1}|s_t, c_t) P(s_{t+1}|s_t, a_t),$$

for all $(x_t, x_{t+1}) \in \mathcal{X} \times \mathcal{X}$.

We arrive at the additional types of agents by directly manipulating their architecture so that the influence the actuators receive are limited, as depicted in Figure 4.6. Here and in the following figures we only display one timestep, t to $t+1$, in order to simplify the depictions.

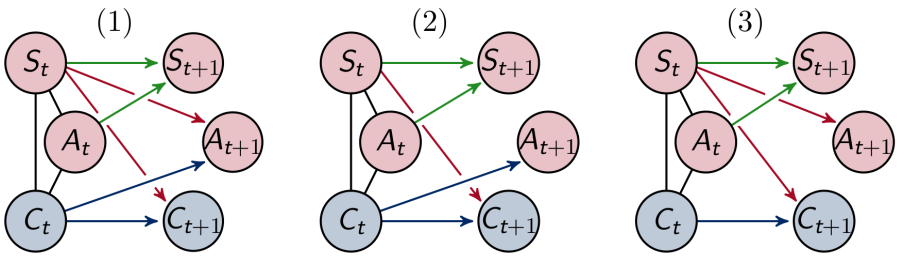


Figure 4.6: The three types of agents: (1) fully coupled agent (2) controller driven agent and (3) reactive control agent.

The first graph (1) of Figure 4.6 depicts the fully coupled agents. In this case the actuators are being influenced by the sensor as well as the controller nodes.

In the middle of Figure 4.6 is the graph depicted that belongs to a class of agents that we call "controller driven". These agents are incapable of sending information directly from the sensor to the actuator nodes. Therefore, all the information has to flow through the controller and gets processed before it reaches the actuators. This leads to the conditional distribution on A_{t+1} being independent of S_t given C_t

$$P(A^i_{t+1}|C_t, S_t) = P(A^i_{t+1}|C_t).$$

The last subclass of agents consists of those for which the controller has no impact on the actuators at all. These agents are called "reactive control" agents because their behavior is only influenced by the direct information flow from the sensors to the actuators, as shown in Figure 4.6 (3). In the case of these reactive agents A_{t+1} is independent of C_t given S_t

$$P(A^i_{t+1}|C_t, S_t) = P(A^i_{t+1}|S_t).$$

Note that both, the controller driven and the reactive control agents, are subclasses of the fully coupled agents. Hence, optimizing the likelihood of success for these agents should not lead to a better outcome compared to the fully coupled agents. However, since we are using the *em*-algorithm that can converge to local minima, this outcome is not guaranteed. In Section 4.6 we observe that the controller driven agents can have a higher likelihood of success compared to the fully coupled agents and we discuss this further in the context of Figure 4.28.

By additionally examining the reactive control and controller driven agents we gain insights in to the importance of the direct connection between the sensors and the actuators as well as between the controller nodes and the actuators. In Section 4.5 we define different information theoretic measures in order to quantify the information flow in the agents fully. Using these measures we are able to observe the differences among the three types of architectures.

4.5 The Measures of the Information Flow

In this section we define twelve different measures for various information flows in our agents. Following the reasoning of the Integrated Information measures, discussed in Chapter 3, these are information theoretic measures that use the KL-divergence to calculate the difference between the original distribution and a split distribution. This split distribution is the one that is closest to the original distribution without having the connection that we want to measure. Hence, this leads to the following definition.

Definition 19 (Measure $\Psi_{\mathcal{M}}$)**.** Let $\mathcal{M} \subset P^\circ(\mathcal{Z} \times \mathcal{C})$ be a set of probability distributions corresponding to a split system. Then we define the measure $\Psi_{\mathcal{M}}$ by minimizing the KL-divergence between $\mathcal{M}$ and the full distribution $\hat{P}$ to quantify the strength of the connections missing in the split system

$$\Psi_{\mathcal{M}} = \inf_{Q \in \mathcal{M}} D_{\mathcal{Z}\times\mathcal{C}}(\hat{P} \parallel Q) = \inf_{Q \in M} \sum_{z, c_{t+2}} \hat{P}(z, c_{t+2}) \log \frac{\hat{P}(z, c_{t+2})}{Q(z, c_{t+2})}.$$

Nearly every discussed measure has a closed form solution and can be written in the form of sums of conditional mutual information terms. We introduce the mutual information, $I(Z_1; Z_2)$, and conditional mutual information, $I(Z_1; Z_2|Z_3)$, in Section 2.2. The conditional mutual information quantifies the connection between Z_1 and Z_2 while Z_3 is fixed. If $I(Z_1; Z_2|Z_3) = 0$, then Z_1 is independent of Z_2 given Z_3. This can also be written as the conditional independence statement $Z_1 \perp\!\!\!\perp Z_2|Z_3$.

Although some of the following measures refer to previously published ones that were originally defined for only one timestep, we introduce them directly tailored to our setting with two timesteps.

The rest of this section is structured as follows. We first define a measure for the total information flow Ψ_{TIF} that is an upper bound for every other measure that we calculate here. We divide the remaining eleven measures into three categories and for each introduced measure we display the corresponding graphical model, if existent, and emphasize the connection quantified by the measure by a dashed arrow. To simplify the depiction the figures only show one timestep. The connections between (Y_{t+1}, Y_{t+2}) are the same as the connections between (Y_t, Y_{t+1}).

First we discuss the measures regarding the controller, which includes the Integrated Information measure. In general, these are the ones that quantify the information flow sent by a controller node. Afterwards we relate them to each other.

Next we focus on measures calculated on the empirical world model, $\tilde{P}(S_{t+1}|S_t, A_t)$. In this context we discuss ways to quantify Morphological Computation and the relationship between the different approaches. The last category contains measures that analyze the information flow coming from the sensors. Thereby we observe the way the sensory information is distributed in the agents.

In addition to the definitions of these measures we also discuss their dynamics based on the results of our experiments. In these experiments we vary the reach of the sensors from 0.5 to 2.75 in steps of 0.25. For each case we took 100 random input distributions and apply the *em*-algorithm. Each time the algorithm performs at least 1000 iteration steps and only stops when the difference between the likelihood of the goal is smaller than $1 * 10^{-5}$. The main results of the experiments are then discussed in Section 4.6.

4.5.1 The Total Information Flow

The measure discussed in this section is called the "total information flow" because there we remove all the connections between the different points in time in the split system. Hence, the set of distributions in the split system can be written as

$$\mathcal{M}_{TIF} = \Big\{ P \in \mathcal{P}^\circ(\mathcal{Z} \times \mathcal{C}) \Big| P(z, c_{t+2}) = P(x_t)P(s_{t+1})P(s_{t+2}) \prod_i P(a^i_{t+1})P(a^i_{t+2}) \prod_j P(c^j_{t+1})P(c^j_{t+2}),\ \forall (z, c_{t+2}) \in \mathcal{Z} \times \mathcal{C} \Big\}$$

Following the Definition 19 the measure for the total information flow results in a sum of mututal information terms

$$\begin{aligned} \Psi_{TIF} &= \inf_{Q \in M} D_{\mathcal{Z}\times\mathcal{C}}(\hat{P} \parallel Q) \\ &= \sum_{\tau \in \{t, t+1\}} \Big(I(S_{\tau+1}; S_\tau, A_\tau) \\ &\quad + I(A_{\tau+1}; S_\tau, C_\tau) + I(C_{\tau+1}; C_\tau, S_\tau) \Big). \end{aligned} \tag{4.3}$$

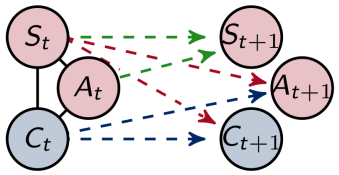


Figure 4.7: Graphical representation of the split model in the case Ψ_{TIF}.

The proof of the result of this minimization is written below. Since the derivation of most of the other measures is very similar, we omit the details in the following paragraphs.

The closed formulation for the Ψ_{TIF} results from the fact the cross-entropies are greater or equal to entropies, which leads to the following inequality

$$
\begin{aligned}
& D_{\mathcal{Z}\times\mathcal{C}}(\hat{P} \parallel Q) \\
&= \sum_{\tau\in\{t,t+1\}} \sum_{x_t, x_{t+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \frac{\hat{P}(x_\tau)\hat{P}(s_{\tau+1}|s_\tau, a_\tau)\prod_i \hat{P}(a^i_{\tau+1}|s_\tau, c_\tau)\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c_\tau)}{Q(x_t)Q(s_{\tau+1})\prod_i Q(a^i_{\tau+1})\prod_j Q(c^j_{\tau+1})} \\
&\geq \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \frac{\hat{P}(x_\tau)\hat{P}(s_{\tau+1}|s_\tau, a_\tau)\prod_i \hat{P}(a^i_{\tau+1}|s_\tau, c_\tau)\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c_\tau)}{\hat{P}(x_\tau)\hat{P}(s_{\tau+1})\prod_i \hat{P}(a^i_{\tau+1})\prod_j \hat{P}(c^j_{\tau+1})} \\
&= \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\hat{P}(s_{\tau+1}|s_\tau, a_\tau)\prod_i \hat{P}(a^i_{t+2}|s_{t+1}, c_{t+1})\prod_j \hat{P}(c^j_{t+1}|s_t, c_t)}{\hat{P}(s_{\tau+1})\prod_i \hat{P}(a^i_{\tau+2})\prod_j \hat{P}(c^j_{\tau+1})} \right) \\
&= \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; S_\tau, A_\tau) + I(A_{\tau+1}; S_\tau, C_\tau) + I(C_{\tau+1}; C_\tau, S_\tau).
\end{aligned}
$$

Now we can take a look at the results in Figure 4.8. This Figure depicts the graphs for the three different types of agents that we introduced in Section 4.4. There we see that for all the agents the total information flow monotonically increases between a sensor length of 1.25 and 2.75. This dynamic is expected since longer sensors detect the environment better. For shorter sensors the agents almost never sense a wall and when they do the likelihood of them dying in the next step is high. Therefore, the longer the sensors are, the more meaningful the information they receive is and the more important the information flows inside the system becomes. We discuss the Ψ_{TIF} between 0.5 and 1.25 in the context of Figure 4.26 further.

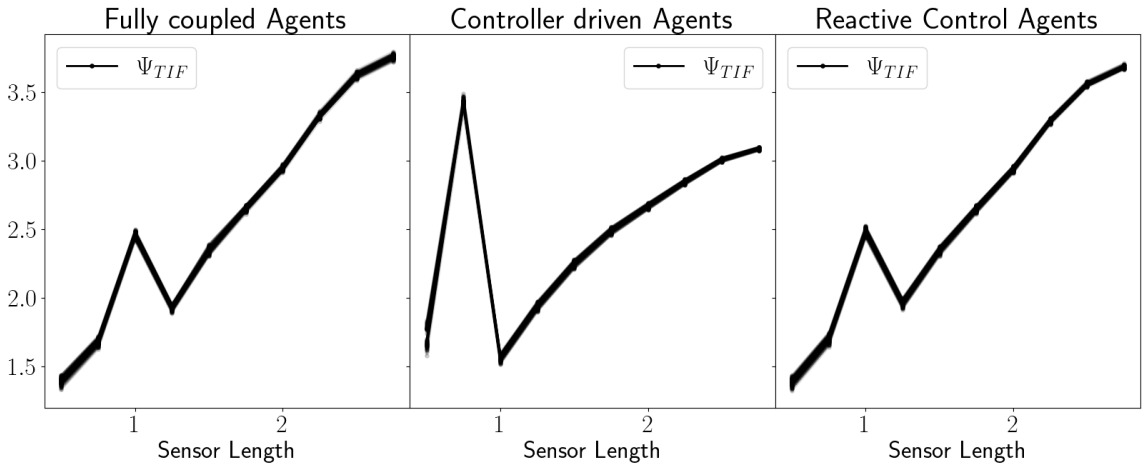


Figure 4.8: The results for the total information flow, Ψ_{TIF}.

4.5.2 Information Flow regarding the Controller

Here in this section we take a close look at the information flows coming from the controller. The corresponding connections are represented by blue arrows. This includes a measure, Φ_T, for Integrated Information.

Integrated Information

The measure Φ_T, ground truth Integrated Information, quantifies how much information is integrated among the controller nodes. It can be seen in the context of the Integrated Information Theory of consciousness, which we discuss in Section 3.1 in detail.

An Integrated Information measure aims at quantifying the strength of the connections among different nodes across different points in time, in other words, the connections that integrate the information.

In Chapter 3 we analyze different Integrated Information measures and come to the conclusion that the ground truth Integrated Information, Φ_T, is the true measure if all of the influences on the controller nodes are known. Since every influence on C_{t+1} is known in our setting, we are able to apply the measure Φ_T

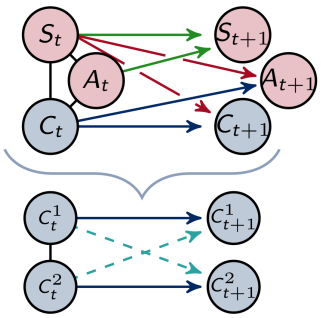


Figure 4.9: Graphical representation of the Integrated Information measure Φ_T.

$$\Phi_T = \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c_\tau)}{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c^j_\tau)} \right)$$
$$= \sum_{\tau\in\{t,t+1\}} \sum_j I(C^j_{\tau+1}; C^{J\setminus\{j\}}_\tau | C^j_\tau, S_\tau).$$

The graphical representation of the corresponding split system is depicted in Figure 4.9. In the definition above $J \setminus \{j\}$ is the set of indices of controller nodes without j. Hence Φ_T measures the connections between C^j_t and $C^{j'}_{j+1}$ with $j, j' \in \{1, 2\}$ and $j \neq j'$.

This definition differs from the one in Proposition 3 slightly because here we calculate the Integrated Information value for two timesteps and because we can specify the exterior influence W_C, that we use in Section 3.6, to be S_τ in the scenario.

In Figure 4.10 we display the different Integrated Information measures, discussed in Chapter 3, for 10 different fully coupled agents. We observe that they all display a similar dynamic, more precisely, decreasing values with an increasing sensor length. Here Φ_T is larger than the other measures, except for Φ_I at a sensor length of 0.5. Every measure apart from Φ_T is calculated only on $\mathcal{C} \times \mathcal{C}$. Hence, in this example the influence from S_t counteracts the existing Integrated Information such that measures that only consider the marginalized distribution are much lower than Φ_T.

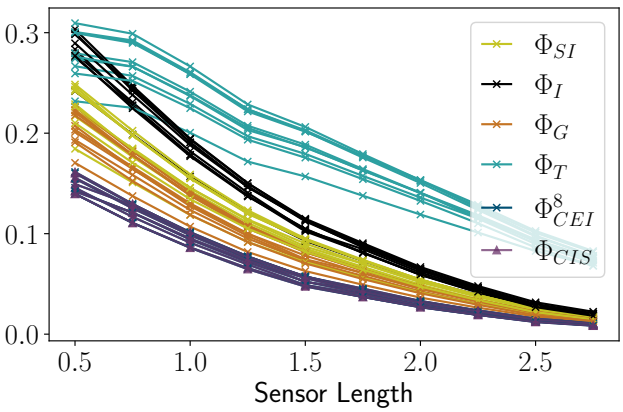


Figure 4.10: The different Integrated Information measures discussed in Chapter 3 for 10 fully coupled agents.

The next measure that we define is an upper bound for the ground truth Integrated Information, as we discuss in the context of Figure 4.13.

Memory

The measure Ψ_M, called "memory", quantifies the total connection between C_t and C_{t+1} given S_t. In this case we remove the entire information flow between the controller nodes for the split system, as depicted in Figure 4.11.

Hence, the measure can be calculated as follows

$$\begin{aligned}\Psi_M &= \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c_\tau)}{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau)} \right) \\ &= \sum_{\tau\in\{t,t+1\}} \sum_j I(C^j_{\tau+1}; C_\tau | S_\tau).\end{aligned}$$

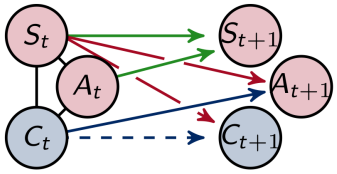


Figure 4.11: Graphical representation of the split model for Ψ_M.

If this value is zero then $C^j_{t+1} \perp\!\!\!\perp C_t | S_t$ and there is no information retained in the controller between different points in time. This would lead to an agent that is similar to the memoryless controller discussed in Section 9.2.2 (3) of [Ay14].

Control

It remains to discuss the information flow from the controller to the actuator nodes. We call the measure quantifying the strength of the impact of the controller on the actuators "control", Ψ_C.

$$\begin{aligned}\Psi_C &= \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\prod_i \hat{P}(a^i_{\tau+1}|s_\tau, c_\tau)}{\prod_i \hat{P}(a^i_{\tau+1}|s_\tau)} \right) \\ &= \sum_{\tau\in\{t,t+1\}} \sum_i I(A^i_{\tau+1}; C_\tau | S_\tau)\end{aligned}$$

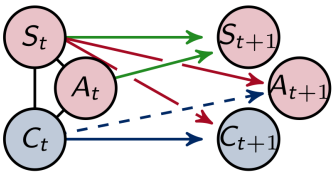


Figure 4.12: Graphic representation of the split model in the case of Ψ_C.

Since we are interested in analyzing the relationship between the internal information flow and the behavior of the agents, this measure plays an important role.

If this measure is zero then the controller has no influence on the actuator state and therefore the movements of the agents.

Relationships among Φ_T, Ψ_M and Ψ_C

Now we discuss the relationship among the measures Φ_T, Ψ_M and Ψ_C. As mentioned earlier, the measure for memory is an upper bound for the Integrated Information measure because

$$\begin{aligned}&\Psi_M - \Phi_T \\ &= \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \left(\log \left(\frac{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c_\tau)}{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau)} \right) - \log \left(\frac{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c_\tau)}{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c^j_\tau)} \right) \right)\end{aligned}$$

This leads to

$$\begin{aligned}\Psi_M - \Phi_T &= \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau, c^j_\tau)}{\prod_j \hat{P}(c^j_{\tau+1}|s_\tau)} \right) \\ &= \sum_{\tau\in\{t,t+1\}} \sum_j I(C^j_{\tau+1}; C^j_\tau | S_\tau)\end{aligned} \tag{4.4}$$

and the conditional mutual information is greater or equal to 0. Similar calculations show that the total information flow, Ψ_{TIF} defined in (4.3), is an upper bound for these measures. More precisely, the relations

$$\begin{aligned}\Phi_T &\leq \sum_{\tau\in\{t,t+1\}} I(C_{\tau+1}; C_\tau, S_\tau) \\ \Psi_M &\leq \sum_{\tau\in\{t,t+1\}} I(C_{\tau+1}; C_\tau, S_\tau) \\ \Psi_C &\leq \sum_{\tau\in\{t,t+1\}} I(A_{\tau+1}; S_\tau, C_\tau)\end{aligned}$$

hold. Therefore we have the following relationships among these measures

$$\Phi_T \leq \Psi_M \leq \Psi_{TIF} \geq \Psi_C.$$

We can observe these relations in Figure 4.13 that shows the results of our experiments for the measures for Integrated Information and memory in the top row and control on the bottom. In addition, we also display the difference between memory and Integrated Information in the upper row.

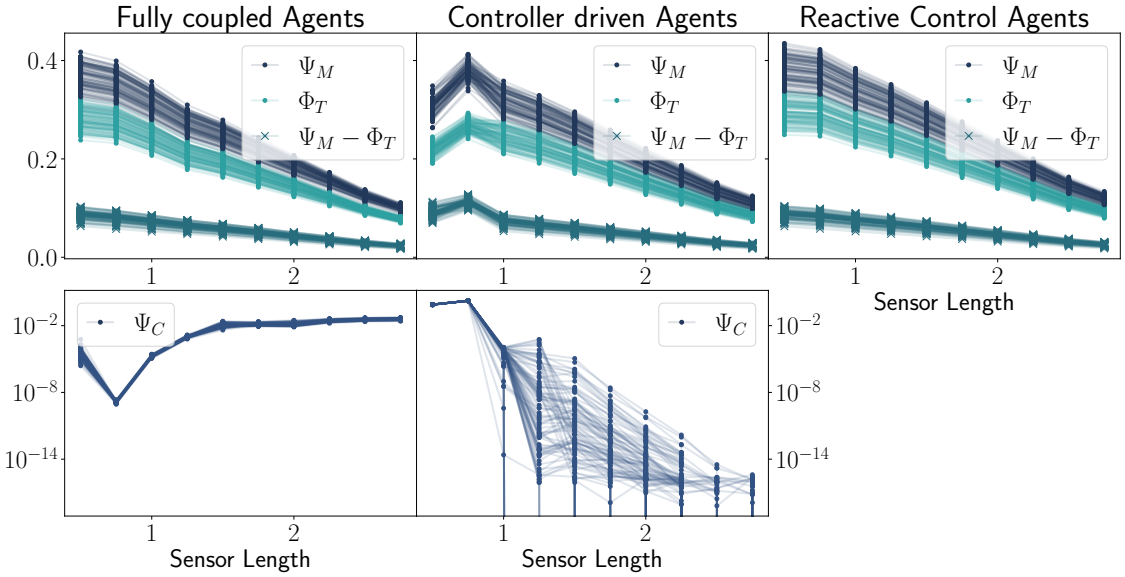


Figure 4.13: The results for the measures for Integrated Information, memory and their difference in the top row and the measure for control in the bottom row.

In this figure we observe that Ψ_M, Φ_T and their difference display a similar decreasing dynamic for all three types of agents while the sensor lengths increase. We discuss the decrease of Φ_T in connection with Morphological Computation in Section 4.6 further.

The difference between Ψ_M and Φ_T results in a sum of conditional mutual information terms, as shown in (4.4). This value is 0 if and only if $C^j_{\tau+1} \perp\!\!\!\perp C^j_\tau | S_\tau$ for all $j \in \{1, 2\}$ and $\tau \in \{t, t+1\}$. In that case the memory only consists of Integrated Information. Here we see that the difference between Ψ_M and Φ_T is low. Hence, a large portion of the information flow inside the controller consists of Integrated Information.

In the bottom row of Figure 4.13 we see the control, Ψ_C, for the fully coupled and controller driven agents. In both cases the values are very low and for reactive agents there is no connection between the controller and the actuators and therefore $\Psi_C = 0$. The fully coupled agents have their lowest control values between a sensor length of 0.5 and 1 and a slightly higher value after that. That means that for short sensors the controller has next to no influence on the behavior of the agent. So, if the sensors are too short, then the agents have to directly act on the information they receive, using reactive control, instead of processing the information in the controller.

However, the controller driven agents are not able to use reactive control. There we see that the control value is higher for the sensor length between 0.5 and 1 compared to the fully coupled agents. Interestingly, for longer sensors this value decreases to almost 0. This means that, although we named these agents "controller driven", they are, in fact, not governed by the controller. We discuss this phenomenon in Section 4.6 in the context of Figure 4.29 further.

4.5.3 Information Flow in the Empirical World Model

This section contains five different measures that we calculate in order to analyze the empirical world model $\tilde{P}(S_{t+1}|S_t, A_t)$. The influence the last actuator and sensor states have on the next sensor state, through the mechanics of the world, are encoded in this distribution. Hence, we discuss measures connected to the concept of Morphological Computation here.

The connections corresponding to the empirical world model are depicted in the figures as green arrows.

Morphological Computation

The concept of Morphological Computation describes the reduction of computational cost for the controller that result from the interaction of the agent's body with its environment. We introduce this concept in more detail in Section 4.1.1.

In [Ghazi-Zahedi13; Ghazi-Zahedi19] the authors define the following measure for Morphological Computation

$$\begin{aligned} \Psi_{MC} &= \sum_{\tau \in \{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\hat{P}(s_{\tau+1}|s_\tau, a_\tau)}{\hat{P}(s_{\tau+1}|a_\tau)} \right) \\ &= \sum_{\tau \in \{t,t+1\}} I(S_{\tau+1}; S_\tau | A_\tau). \end{aligned}$$

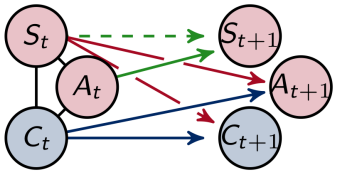


Figure 4.14: Graphical representation of the split model in the case of Ψ_{MC}.

It quantifies the strength of the influence of the past sensory input on the next sensory input given the last action. The graphical representation associated with the split system of Ψ_{MC} is depicted in Figure 4.14.

In [Ghazi-Zahedi13; Ghazi-Zahedi19] this measure is called the "Associative measure of the positive effect of the world", $ASOC_W$ and in [Ghazi-Zahedi19] the author compares

different candidate measures for Morphological Computation numerically. He concludes in the Chapter 4.9 that the measure following the approach of Ψ_{MC}, but defined directly on the world states, has advantages over other formulations and is therefore the recommended one. However, we do not have a direct access to the world states but in Proposition 9 we show, under a technical assumption, that marginalizing over the world leads to the empirical world model. Hence, we consider Ψ_{MC} to be the measure of Morphological Computation.

Another candidate for a Morphological Computation measure is called "action effect".

Action Effect

Here we quantify the effect the actuator has on the next sensory state by calculating

$$\Psi_{AE} = \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log\left(\frac{\hat{P}(s_{\tau+1}|s_\tau, a_\tau)}{\hat{P}(s_{\tau+1}|s_\tau)}\right)$$
$$= \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; A_\tau|S_\tau)$$

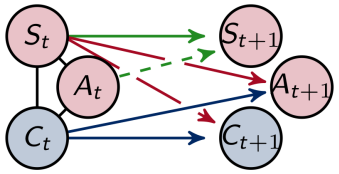


Figure 4.15: Graphical representation of the split model in the case of Ψ_{AE}.

and we call this measure "action effect". This measures the amount of influence an agent has on its environment. Hence in [Ghazi-Zahedi13] this measure was normalized and inverted in order to quantify Morphological Computation. The differences between this approach and Ψ_{MC} are further discussed in Section 4.9 in [Ghazi-Zahedi19].

This measure is called "causal action influence" in [Seitzer21], where the authors integrate this measure into reinforcement learning algorithms in order to improve the learning efficiency.

Predictive Information

The measures discussed here assess how useful information from the past is to predict the next sensory state. This is done by calculating the mutual information between the past and present sensory state. In Section 3 of [Grassberger86] this was introduced as a measure for the complexity of ensembles of patterns and named "effective measure complexity". Bialek et al. further analyze this measure in [Bialek01] and name it "predictive information". The predictive information in the sensor space of robots was used in [Ay08a] as an objective function for self-organization and there it proves effective in a simulated chain of passively coupled two wheeled robots in a maze.

The first measure for predictive information is restricted to the sensor states S_{t+1}, S_t and therefore called "sensory prediction"

Figure 4.16: Graphical representation of the split model in the case of Ψ_{SP}.

$$\Psi_{SP} = \sum_{\tau\in\{t,t+1\}} \sum_{s_\tau, s_{\tau+1}} \log\left(\frac{\hat{P}(s_{\tau+1}|s_\tau)}{\hat{P}(s_{\tau+1})}\right)$$
$$= \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; S_\tau).$$

This reflects the scenario in which we are not able to include the action in our world model and only rely on the sensor states.

However, in our case we are able to include the last actuator value to calculate how predictable the next sensory input is given the knowledge of the last sensor states and the initiated action.

This measure is called "environment predictability".

$$\Psi_{EP} = \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log \left(\frac{\hat{P}(s_{\tau+1}|s_\tau, a_\tau)}{\hat{P}(s_{\tau+1})} \right)$$
$$= \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; S_\tau, A_\tau).$$

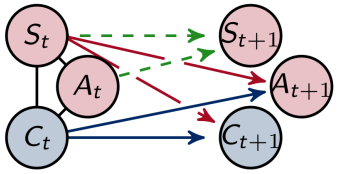


Figure 4.17: Graphical representation of the split model in the case of Ψ_{EP}.

If this measure is zero then the next sensory input does not depend on the past sensor state or the past action. In this case we would not be able to optimize the behavior of the agent since it has no impact on its survival.

Synergistic Information

The last measure that we discuss in the context of the empirical world model is conceptually different from the other measures. More precisely, for this measure the split model is not a graphical model and the minimization in Definition 19 has no closed form solution.

In this case the split model $\mathcal{M}_{Syn}$ is not simply missing one or more connections, instead its elements have no synergistic interactions among S_t and A_t on S_{t+1}. This approach was proposed in [Ghazi-Zahedi17a] as a measure for Morphological Computation and is called synergistic information Ψ_{Syn}.

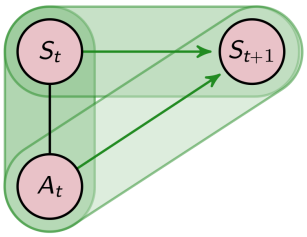


Figure 4.18: Sketch of the split system of Ψ_{Syn}.

The name "synergistic information" suggests a connection to the problem of finding a correct information decomposition of random variables. In that context one aims at decomposing the information that a set of variables holds about a target variable into separate, non-negative terms, namely into redundant, synergistic and unique information. There exist different definitions of these components, see for example [Williams10; Harder; Ay20; Kolchinsky22]. In [Ghazi-Zahedi19] in Section 3.5 the author discusses the unique information that results from taking the above approach to synergistic information.

Removing only the synergistic interactions does not result in a graphical model but in a generalization called hierarchical model, as described in Section 2.4.

A hierarchical model is an exponential family defined by a simplicial complex, for more details see for example Section 2.9 in [Ay17]. The facets of the simplicial complex in our context are given by the following three sets of random vectors $\{A_t, S_t\}$, $\{A_t, S_{t+1}\}$ and $\{S_t, S_{t+1}\}$. These sets are highlighted in the sketch in Figure 4.18 in different shades of green. Hence the spanning functions f of the vector space F in Definition 4 only depend on one of these three sets each.

We remove the three-way interaction of (S_t, A_t) on S_{t+1} by using the iterative scaling algorithm to find a distribution $P \in \mathcal{M}_{Syn}$ that fixes the two-way interactions among the three parts A_t, S_t and S_{t+1} to the two-way interaction of the initial distribution $\hat{P}$, meaning

$$P(A_t, S_t) = \hat{P}(A_t, S_t), \;\; P(A_t, S_{t+1}) = \hat{P}(A_t, S_{t+1}) \text{ and } P(S_t, S_{t+1}) = \hat{P}(S_t, S_{t+1}). \quad (4.5)$$

The iterative scaling algorithm was introduced earlier in Section 2.5.2 and it works by iteratively performing e-projections to linear families defined by the constraints above.

Using the iterative scaling algorithm provides us with an additional interpretation of this measure and its split distribution. The result in Theorem 2.5.2 shows that the m-projection

to $\mathcal{M}_{Syn}$ is equivalent to finding the maximum entropy solution. Therefore, the iterative scaling algorithm converges to a distribution that satisfies the constraints in (4.5) without making any further assumptions on the structure of the split distribution.

Note that in the constraints in (4.5) we only use unconditioned marginal distributions, however, in the original graph the connections between S_t and S_{t+1} as well as between A_t and S_{t+1} are directed. Nonetheless, it is easy to see that the iterative scaling algorithm fixes the required conditional distributions as well, because the limit point $P^\star$ of the algorithm has the same marginals as $\hat{P}$. Hence, we also have the following equalities

$$P^\star(S_{t+1}|S_t) = \hat{P}(S_{t+1}|S_t) \text{ and } P^\star(S_{t+1}|A_t) = \hat{P}(S_{t+1}|A_t).$$

Furthermore, we can directly observe that the iterative scaling algorithm fixes the conditioned distribution in the following way. Let ℓ be one step in the iterative scaling algorithm where P^ℓ was projected to the linear family defined by (A_t, S_t). Hence $P^\ell(A_t, S_t) = \hat{P}(A_t, S_t)$ and now we project to one of the two other linear sets. Without loss of generality we choose the one connected to (S_{t+1}, A_t), then the next step can be written as follows

$$\begin{aligned} P^{\ell+1}(s_{t+1}, s_t, a_t) &= P^\ell(s_{t+1}, s_t, a_t)\frac{\hat{P}(s_{t+1}, a_t)}{P^\ell(s_{t+1}, a_t)} \\ &= P^\ell(s_{t+1}, s_t, a_t)\frac{\hat{P}(s_{t+1}|a_t)}{P^\ell(s_{t+1}|a_t)} \\ &= P^\ell(s_t|s_{t+1}, a_t)\hat{P}(s_{t+1}|a_t)P^\ell(a_t), \end{aligned}$$

for all $(s_{t+1}, s_t, a_t) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$. So we can write the initial e-projection in a way that fixes the conditional distribution. This only works because $P^\ell(A_t) = \hat{P}(A_t)$. However, the distribution $P^{\ell+1}(S_t)$ is in general not equal to $\hat{P}(S_t)$. Therefore we have to project to the linear set defined by (S_t, A_t) in every second step. In [Brown93] the authors show that the convergence of the iterative scaling algorithm is order-independent as long as each linear family is visited infinitely often. Hence, both approaches, fixing the joint or the conditional distributions, are equivalent in this case.

For a variant of the iterative scaling algorithm, which is called the generalized iterative scaling algorithm, where all of the projections are performed at once there exists a version where the constraints are directly defined via conditional distributions. This variation of the algorithm was discussed in for example [Goodman02].

Relationships among $\Psi_{MC}, \Psi_{Syn}, \Psi_{EP}, \Psi_{AE}$ and Ψ_{SP}

Here we relate the different measures on the empirical world model to each other. We start with the environment predictability Ψ_{EP}. In that case all the connections between S_t, A_t and S_{t+1} are removed and therefore this is an upper bound for all the other measures on the empirical world model.

Additionally, it is easy to see from the definition of Ψ_{EP} and Equality (4.3) that $\Psi_{EP} \leq \Psi_{TIF}$.

The synergistic measure Ψ_{Syn} is defined by a split system in which only the three-way interactions of S_t, A_t on S_{t+1} are removed. The split systems of Ψ_{MC} and Ψ_{AE} also do not contain these synergistic interactions because in each case one whole connection from S_t or A_t to S_{t+1} is missing. Therefore, these split systems are subsets of the split

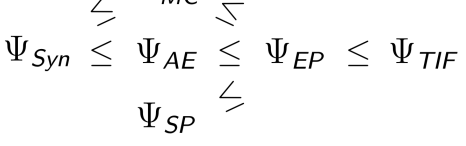


Figure 4.19: Relationships among the measures calculated on the empirical world model.

system of Ψ_{Syn}. This leads to Ψ_{MC} and Ψ_{AE} being greater or equal to Ψ_{Syn}. The connections among the measures are summarized in Figure 4.19.

Furthermore, the predictability of the environment decomposes into the action effect and the prediction of the sensors, as shown below.

$$\begin{aligned}\Psi_{AE} + \Psi_{SP} = &\sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log\left(\frac{\hat{P}(s_{\tau+1}|s_\tau, a_\tau)}{\hat{P}(s_{\tau+1}|s_\tau)}\right) \\ &+ \sum_{\tau\in\{t,t+1\}} \sum_{s_\tau, s_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log\left(\frac{\hat{P}(s_{\tau+1}|s_\tau)}{\hat{P}(s_{\tau+1})}\right)\end{aligned}$$

This leads to

$$\begin{aligned}\Psi_{AE} + \Psi_{SP} = &\sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} \hat{P}(x_\tau, x_{\tau+1}) \log\left(\frac{\hat{P}(s_{\tau+1}|s_\tau, a_\tau)}{\hat{P}(s_{\tau+1})}\right) \\ = &\ \Psi_{EP}.\end{aligned} \tag{4.6}$$

In Figure 4.20 the results for the measures that are calculated on the empirical world model are depicted. In the upper row we see the dynamics of Ψ_{EP}, Ψ_{SP} and Ψ_{MC}, which are very similar.

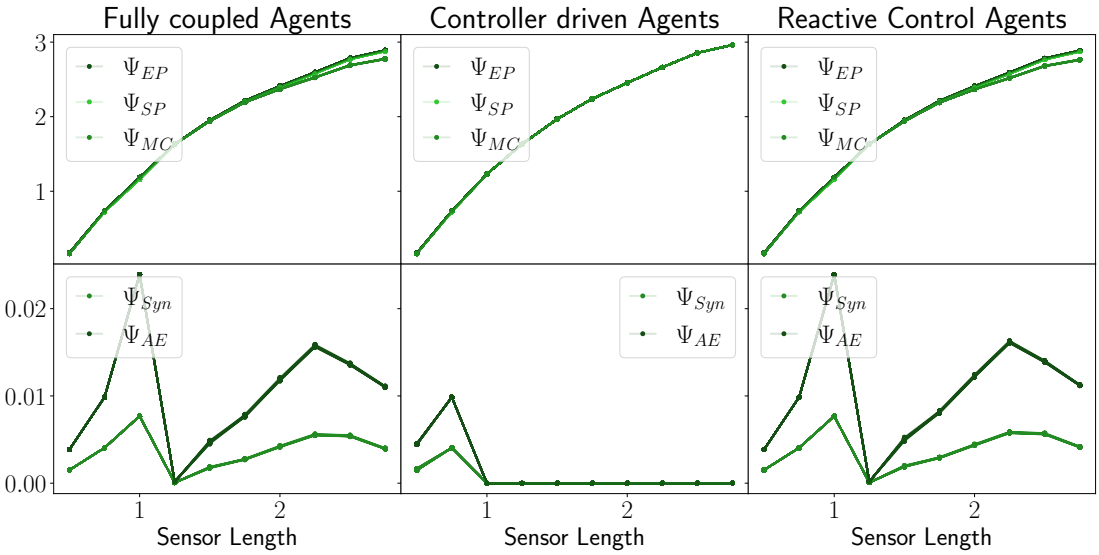


Figure 4.20: Results of the different measures calculated on the empirical world model.

Especially in the case of the controller driven agents there is next to no difference between these three measures. Considering the equation (4.6) this means that the actuator states have close to no influence on the next sensory state. This is confirmed by the results in the bottom row of Figure 4.20. There we observe that the measure for the action effect, Ψ_{AE}, and the synergistic information, Ψ_{Syn}, are both low compared to Ψ_{SP} especially in the case of the controller driven agents. We discuss the implications of this for the behavior of the agents in Section 4.6 in the context of Figure 4.29.

Note that the variances in the results for the different initial distributions are barely visible here in any of these measures. The reason for that is that the empirical world model is sampled and therefore results in the same distribution depending on the sensor length. The *em*-algorithm merely influences a part of the distribution in time t, namely $P(A_t)$. Therefore we are only able to observe very small variations in the results. The value of the Ψ_{Syn}, for instance, lies for the reactive control agents in case of a sensor length of 2.25 between 0.0057 and 0.0059. Hence, these differences are too small to be clearly visible in Figure 4.20.

4.5.4 Flow of the Sensory Information inside the Agents

The last three measures that we discuss quantify the flow of the information from the sensors to the actuator or controller nodes. The respective connections are depicted by red arrows in the figures. First, we introduce two measures on the connection between S_t and A_{t+1} and afterwards we look at a measure for the information flow going from the sensor to the controller nodes.

Reactive Control

The measure named "reactive control", Ψ_R, describes the influence of the direct stimuli response, meaning of unprocessed information that is send directly to the actuator nodes. In the corresponding split distribution the connection between S_t and A_{t+1} is removed, as depicted Figure 4.21. This results in the measure

$$\begin{aligned}\Psi_R &= \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau, x_{\tau+1}} P(x_\tau, x_{\tau+1}) \log \left(\frac{\prod_i P(a^i_{\tau+1}|s_\tau, c_\tau)}{\prod_i P(a^i_{\tau+1}|c_\tau)} \right) \\ &= \sum_{\tau\in\{t,t+1\}} \sum_i I(A^i_{\tau+1}; S_\tau | C_\tau).\end{aligned}$$

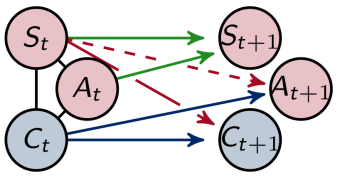


Figure 4.21: Graphic representation of the split model in the case of Ψ_R.

Note that the controller driven agents lack this connection and therefore always have $\Psi_R = 0$.

Multisensory Integration

Additionally, we calculate a second measure regarding this connection called "multisensory integration", Ψ_{MSI}. There we want to quantify how important the integration of data from the different sensor nodes in the actuator nodes is.

Compared to the Integrated Information measures we do not have the same substrates in the two points in time here and therefore no logical reason to identify A^1_{t+1} with S^1_t rather than S^2_t. Hence, we calculate two values with two different split systems and then take the minimum of the result. A sketch of this approach can be found in Figure 4.22.

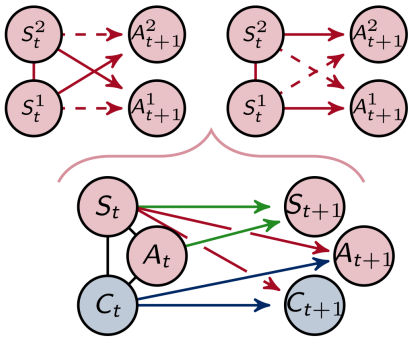


Figure 4.22: Sketch of the split model in the case of the multisensory integration, Ψ_{MSI}.

In the first split system A^1_{t+1} only depends on S^2_t and A^2_{t+1} depends on S^1_t. This is depicted on the top left in Figure 4.22. Additionally, in the second case we use a split system in which A^1_{t+1} depends on S^1_t and A^2_{t+1} depends on S^2_t.

This results in the following measure

$$\Psi_{MSI} = \sum_{\tau\in\{t,t+1\}} min\left\{ \sum_{x_\tau,x_{\tau+1}} P(x_\tau,x_{\tau+1}) \log\left(\frac{P(a^1_{\tau+1}|s_\tau,c_\tau)P(a^2_{\tau+1}|s_\tau,c_\tau)}{P(a^1_{\tau+1}|s^2_\tau,c_\tau)P(a^2_{\tau+1}|s^1_\tau,c_\tau)}\right), \right.$$
$$\left. \sum_{x_\tau,x_{\tau+1}} P(x_\tau,x_{\tau+1}) \log\left(\frac{P(a^1_{\tau+1}|s_\tau,c_\tau)P(a^2_{\tau+1}|s_\tau,c_\tau)}{P(a^1_{\tau+1}|s^1_\tau,c_\tau)P(a^2_{\tau+1}|s^2_\tau,c_\tau)}\right)\right\}$$
$$= \sum_{\tau\in\{t,t+1\}} min\left\{I(A^1_{\tau+1};S^1_\tau|C_\tau,S^2_\tau) + I(A^2_{\tau+1};S^2_\tau|C_\tau,S^1_\tau),\right.$$
$$\left. I(A^1_{\tau+1};S^2_\tau|C_\tau,S^1_\tau) + I(A^2_{\tau+1};S^1_\tau|C_\tau,S^2_\tau)\right\}$$

We sum over the minimum of each timestep respectively. If this measure is 0 then the information from the two sensor nodes do not need to get combined in the actuator nodes.

Sensory Information

The last measure that we discuss here is called "sensory information". It calculates the strength of the information flow from the sensor to the controller nodes, as depicted in Figure 4.23, and it is calculated as follows

$$\Psi_{SI} = \sum_{\tau\in\{t,t+1\}} \sum_{x_\tau,x_{\tau+1}} P(x_\tau,x_{\tau+1}) \log\left(\frac{\prod_j P(c^j_{\tau+1}|s_\tau,c_\tau)}{\prod_j P(c^j_{\tau+1}|c_\tau)}\right)$$
$$= \sum_{\tau\in\{t,t+1\}} \sum_j I(C^j_{\tau+1};S_\tau|C_\tau).$$

The smaller this value is, the more likely it is that the controller converged to a general strategy and performs this blindly without including the information from the sensors. In order for the controller to send meaningful commands to the actuators these commands should be based on the information received from the sensors. Therefore this measure is important for judging the relevance of the Integrated Information. We discuss this further in Section 4.6 in connection with the effective information integration in Definition 20.

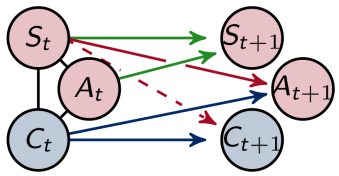


Figure 4.23: Graphic representation of the split model in the case of Ψ_{SI}.

Relationships among Ψ_R, Ψ_{MSI} and Ψ_{SI}

In this section we discuss the relationships among Ψ_R, Ψ_{MSI} and Ψ_{SI}. The two measures Ψ_R and Ψ_{MSI} are both calculated with respect to the connection between the sensors and the actuators. There Ψ_R removes the connection completely while Ψ_{MSI} considers a split system with only a partly removed information flow. Therefore

$$\Psi_{MSI} \leq \Psi_R \leq \Psi_{TIF}.$$

The last inequality is easy to see, since Ψ_R is smaller or equal to the second summand in (4.3). Similarly, $\Psi_{SI} \leq \Psi_{TIF}$, considering the last summand in the definition of the total information flow.

Figure 4.24 depicts the results for the three different types of agents for the sensory information in the top row and reactive control, multisensory integration and their difference, in the case of the fully coupled and reactive control agents, in the bottom row.

The measure Ψ_{SI} quantifies how important the information flow from the sensors to the controller is. For a length below 1 the sensors are too short and above approximately 2 too long to carry information that is valuable for the controller. Therefore, in the cases of the fully coupled and reactive control agents this value increases with the length of the sensors until 1.5 after which it decreases again.

The controller driven agents depict roughly the same behavior for sensors longer than 1. For a sensor length of 0.5 and especially 0.75, the sensory information is much more important compared to the other types of agents. We discuss this in connection with other information flows in the context of Figure 4.29.

The measure for reactive control is shown in the second row of Figure 4.24. For the controller driven agents this measure is 0 per definition. In the case of short sensors the information needs to get passed directly to the actuators and therefore the corresponding information flow is important, Ψ_R is high. Between 1.25 and 2 the sensors are long enough that the information of touching a wall can be send to the controller nodes, which then send that information to the actuators. Above approximately a sensor length of 2 the sensors are so long that them touching a wall is not valuable information. Instead, the agents react to not touching a wall. This leads once again to an increase in the importance of reactive control.

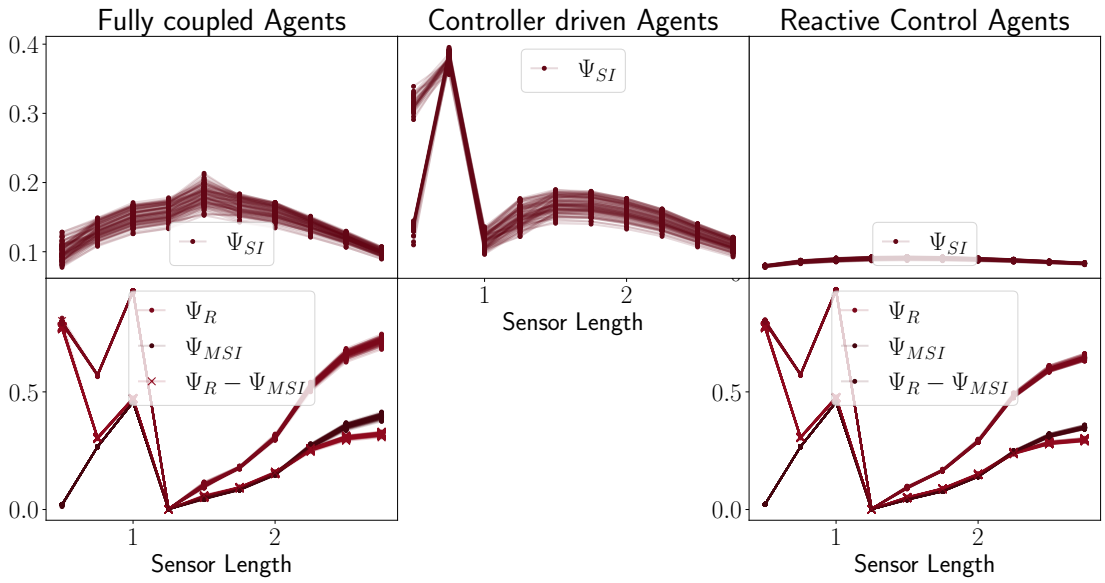


Figure 4.24: The measures regarding the information flowing from the sensors to the actuator or controller nodes.

The second measure calculated on the connection between the sensors and the actuator nodes is the multisensory integration Ψ_{MSI}. In the bottom row of Figure 4.24 we observe that the dynamics of this measure is very similar to Ψ_R, only with an overall lower value, except in the case of sensors with a reach of 0.5. Then the multisensory integration is close to 0, hence, it is not important that the information from both sensors gets combined but the information that one of the sensors touches a wall suffices. In this case the sensors are so short that if one sensor detects a wall the agent has to react immediately, regardless of the value of the other sensor.

Additionally, we discuss the difference between multisensory integration and reactive control.

$$
\begin{aligned}
&\Psi_R - \Psi_{MSI} \\
&= \sum_{\tau \in \{t,t+1\}} max \Bigg\{ \sum_i I(A^i_{\tau+1}; S_\tau | C_\tau) - \left(I(A^1_{\tau+1}; S^1_\tau | C_\tau, S^2_\tau) + I(A^2_{\tau+1}; S^2_\tau | C_\tau, S^1_\tau) \right), \\
&\qquad\qquad \sum_i I(A^i_{\tau+1}; S_\tau | C_\tau) - \left(I(A^1_{\tau+1}; S^2_\tau | C_\tau, S^1_\tau) + I(A^2_{\tau+1}; S^1_\tau | C_\tau, S^2_\tau) \right) \Bigg\} \\
&= \sum_{\tau \in \{t,t+1\}} max \left\{ I(A^1_{\tau+1}; S^2_\tau | C_\tau) + I(A^2_{\tau+1}; S^1_\tau | C_\tau),\ I(A^1_{\tau+1}; S^1_\tau | C_\tau) + I(A^2_{\tau+1}; S^2_\tau | C_\tau) \right\}
\end{aligned}
$$

Hence, if this difference is zero, then all the conditional mutual information terms $I(A^j_{\tau+1}; S^{j'}_\tau | C_\tau)$ are zero for all $j, j' \in \{1, 2\}$. Then the actuator nodes do not depend on one sensory node, given the controller nodes. In this case the reactive control Ψ_R only consists of the multisensory integration, hence in that case the information from both sensors are combined in the actuator states. In Figure 4.24 we see that such a situation exists only for a sensor length of 1.25 where reactive control and multisensory information are both close to zero.

Except for a sensor length of 0.5 the difference between Ψ_R and Ψ_{MSI} is very similar to Ψ_{MSI}. Hence combining the information from both sensors is about as important as the influence from one sensor on one actuator node given the controller nodes. As mentioned above, for a sensor length of 0.5 the information from one of the sensors is sufficient for the actuators.

This concludes the introduction of the various measures that we apply to the information flow in the agents. In the next section we draw conclusions from the interactions among the different types of measures. This leads to the introduction of the "effective Information Integration".

4.6 Results of the Experiments and Effective Information Integration

In this section we combine the results of all the different types of measures discussed in the previous sections. This allows us to reach conclusions about the importance of the various information flows for the behavior of the agents.

As stated in Section 5.5, we apply the *em*-algorithm to 100 random initial distributions for every type of agent and each sensor length between 0.5 and 2.75 with a step size of 0.25. Afterwards, the measures are calculated for the distribution that the *em*-algorithm reaches after at least 1000 steps when the difference between the likelihoods of reaching the goal between two consecutive steps is smaller than $1 * 10^{-5}$.

Before we take a look at these measures we first discuss how successful the agents perform the task. This can be characterized by the probability of not touching the wall after the next two movements, more precisely by $P(g_1 = 1)$. We calculate this probability by using the *em*-algorithm and therefore this result is purely theoretical. It does not necessarily reflect how well an agent would actually perform inside the racetrack, however, this value allows us to compare the result for the different sensor lengths and types of agents.

The graphs in Figure 4.25 depict this probability of reaching the goal. Here we denote the probability distributions corresponding to the fully coupled, controller driven and reactive control agents by P_1, P_2 and P_3 respectively. Although we use 100 random initial

distributions there is little variation in the resulting probability in each case. We observe that all three types of agents perform best between a sensor length of 1.25 and 2.25. This indicates that the information from the sensors is not as useful to assure the survival of the agents if the sensors are longer than 2.25 or shorter than 1.25.

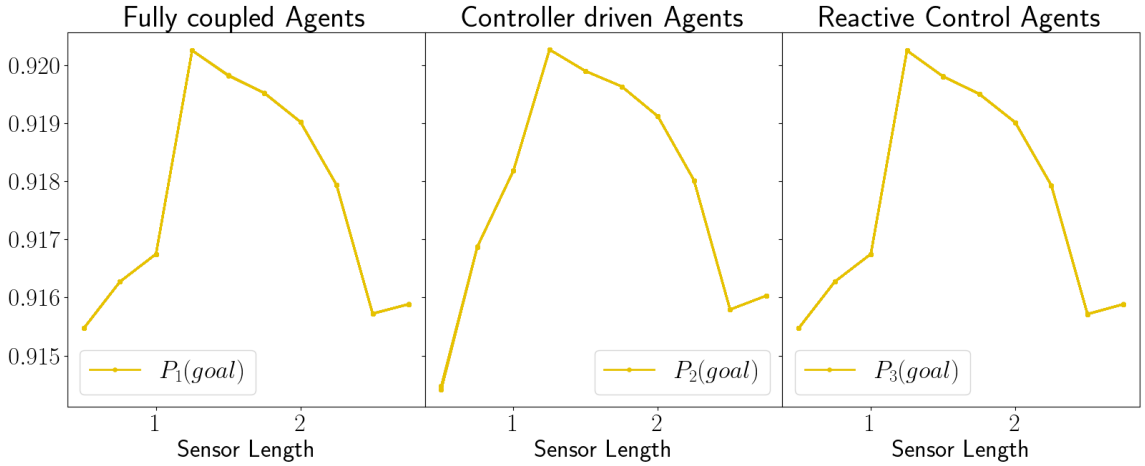


Figure 4.25: The probability of reaching the goal for the fully coupled, controller driven and reactive control agents.

The difference between the results for the fully coupled and the reactive control agents are not visible in this depiction but we take a closer look at it in the context of Figure 4.28. However, there is a visible difference between the controller driven agents and the other two. In the case of sensors with a length of 0.5 the controller driven agents perform worse than the other two types of agents. The reason for this was already discussed in the context of the measure for reactive control Ψ_R. When the sensors are that small the agents have to react directly to a detected wall and the information has to be send immediately from the sensors to the actuators. The controller driven agents are not able to utilize this connection and therefore their performance is worse compared to the fully coupled and reactive control agents.

It is interesting to note that the same caveat, that the controller driven agents are not able to use reactive control, results in an advantage for the sensor length of 0.75 and 1. For these sensor lengths the controller driven agents perform better compared to the fully coupled or reactive control agents. Note that this is only possible because the *em*-algorithm converges to a local optimum. Since the controller driven agents are a subclass of the fully coupled ones, the fully coupled agents would have to perform best otherwise, as discussed in Section 4.4. In order to understand the behavior of the controller driven agents for the sensor lengths 0.75 and 1 we further analyze the different information flows in the context of Figure 4.29.

Now we discuss the first introduced measure, the total information flow Ψ_{TIF}, which decomposes into a sum of three types of mutual information terms, shown in Section 4.5.1. Taking only the first part into account results in a different measure, called environment predictability and introduced in Section 4.5.3, Ψ_{EP}. It calculates how dependent the next sensory state is on the last sensor and actuator states. Hence, in order to simplify the notation we name the remaining components of the total information flow "controller

dependence", Ψ_{CD}, and "actuator dependence", Ψ_{AD}.

$$\Psi_{CD} = \sum_{\tau \in \{t,t+1\}} I(C_{\tau+1}; C_\tau, S_\tau)$$
$$\Psi_{AD} = \sum_{\tau \in \{t,t+1\}} I(A_{\tau+1}; C_\tau, S_\tau)$$

Additionally, we normalize each of these measures by dividing them by Ψ_{TIF}. This allows us to compare the importance of the information flows to the sensors, controller or actuator nodes in proportion to the total information flow. The sum of these normalized measures results in 1.

In Figure 4.26 we depict these three normalized measures. We observe that the results for the fully coupled and reactive control agents look very similar. For these two types of agents the environment predictability is the largest portion of the total information flow for sensors larger than 0.5. This value has its highest point at a sensor length of 1.25. There the information flow towards the sensors constitutes around 85% of the total information flow. In the context of Figure 4.20 we discuss that the information flow between the sensors makes up for most of the value of Ψ_{EP}. This reflects the nature of the experiment. If the agents in the racetrack do not detect a wall, then it is highly likely that they will not detect a wall in the next point in time regardless of their action.

The opposite behavior can be observed for Ψ_{AD}, which assesses the proportional impact of the information flowing to the actuators. For the fully coupled and reactive control agents the minimum is at 1.25. The measures for reactive control, Ψ_R, and control, Ψ_C, each quantify a part of Ψ_{AD}. The higher values between a sensor length of 0.5 and 1.25 can also be observed for Ψ_R, depicted in Figure 4.24, whereas the dynamics of Ψ_{AD} between a sensor length of 1.25 and 2.75 is similar to Ψ_R and Ψ_C.

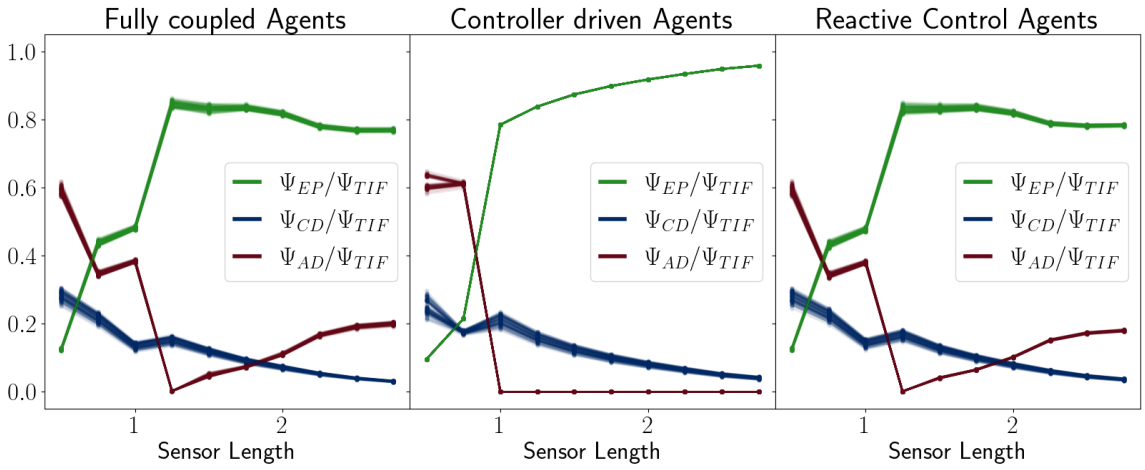


Figure 4.26: The values for Ψ_{EP}, Ψ_{CD} and Ψ_{AD} divided by Ψ_{TIF}.

The controller driven agents lack the ability to use reactive control and the proportional impact of the information flow to the actuators is almost zero in the case of sensors larger than 0.75. In these cases the algorithm converged to a probability distribution for the actuator states that is almost independent of the sensory and controller states. We discuss that in more detail in the context of Figure 4.29.

Since Ψ_{AD} has next to no impact in this case, the other two normalized measures, namely environmental predictability and controller dependence, have to display an antagonistic dynamic. Here Ψ_{EP}/Ψ_{TIF} is monotonically increasing while Ψ_{CD}/Ψ_{TIF} decreases with growing sensors. The importance of the information flow towards the controller nodes is similar across the different types of agents.

The value of Ψ_{CD} consists partly of the Integrated Information measures Φ_T. Similarly, a large share of Ψ_{EP} is Morphological Computation, Ψ_{MC}, as can be seen in Figure 4.20. In the introduction, in Section 4.1, we hypothesize that Morphological Computation and Integrated Information behave asymmetrically. Hence, we now take a closer look at this relationship by analyzing the true, not normalized, results of these measures.

The results for Integrated Information, Φ_T, and Morphological Computation, Ψ_{MC}, measures are depicted in Figure 4.27. Because of the close similarities of the results between the different types of agents the values for Φ_T is depicted for all of the three types on the top and all the values for Ψ_{MC} are shown on the bottom of Figure 4.27.

We observe that Φ_T monotonically decreases as the sensors become larger. Hence, the shorter the sensors are, the more the information gets integrated among the different controller nodes.

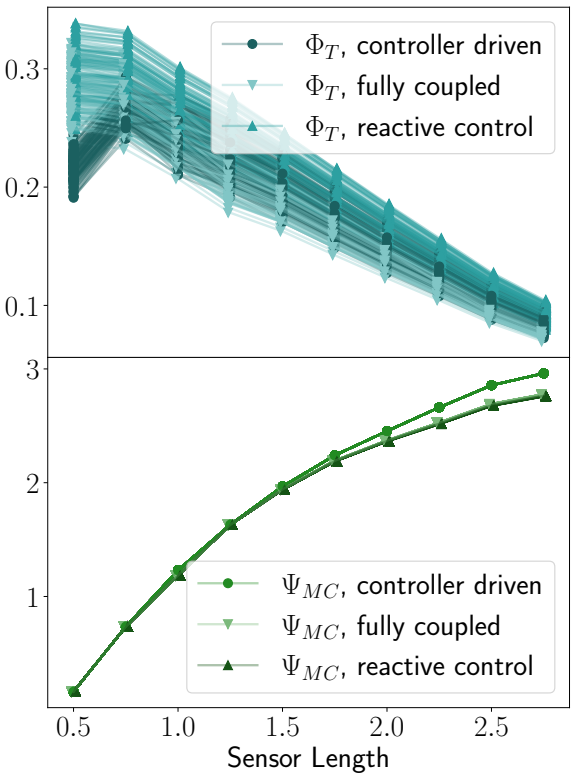


Figure 4.27: Comparison between Integrated information Φ_T and Morphological Computation Ψ_{MC}.

Directly below Φ_T the measure Ψ_{MC} exhibits the opposite dynamic. For all three types of agents the Morphological Computation increases with the length of the sensors. This measure quantifies the influence of the past sensory input on the next sensory input given the action. Hence, it assesses the information flowing through the environment. Starting with a sensor length of 1.75 the controller driven agents have a higher Morphological Computation compared to the rest. As discussed in the context of Figure 4.20, this means that $P(A_t)$ converged to a distribution that facilitates Morphological Computation.

Taking the perspective of the agent Ψ_{MC} describes the extrinsic information flow, whereas Φ_T only depends on the controller nodes and therefore quantifies the intrinsic information flow. So these measures exhibit an antagonistic relationship between the outside and the inside of the agent, meaning between Morphological Computation and Integrated Information.

This relationship can still be observed in the case of the reactive control agents even though the controller has no influence on the actuators.

Note that the connection from the past controller nodes to the next one does not get modified by the *em*-algorithm in this case. For the reactive control agents the *e*-projection

does not alter the distribution $P(C_{t+1}|C_t, S_t)$ because $A_{t+1} \perp\!\!\!\perp C_t|S_t$ and therefore the goal does not depend on the state of the controller. Hence, the decrease with the length of the sensors is simply caused by

$$P(s_t, c_t) = \sum_{a_t} P(c_t|s_t, a_t)\tilde{P}(s_t|a_t)P(a_t).$$

Therefore, we can conclude that while the measures for Morphological Computation and Integrated Information behave antagonistically this does not allow us to draw conclusions about the importance of the Integrated Information for the behavior of the agent.

The same situation can be observed in the case of the fully coupled agents. Although Φ_T has its highest values for shorter sensor lengths, the Integrated Information does not have a significant impact on the behavior of the agents because the measure for control is close to 0. More precisely, we see in Figure 4.13 that for a sensor length of 0.75 the value for control lies between 10^{-8} and 10^{-10}.

In Figure 4.28 on the top we compare the measures for the sensory information, Ψ_{SI}, and control, Ψ_C. Between a sensor length of 1.5 and 2.5 both measures are relatively high. Here we judge whether a measure is high or low by considering its values for all the sensor length and not by comparing it to other measures, because they are on vastly different scales.

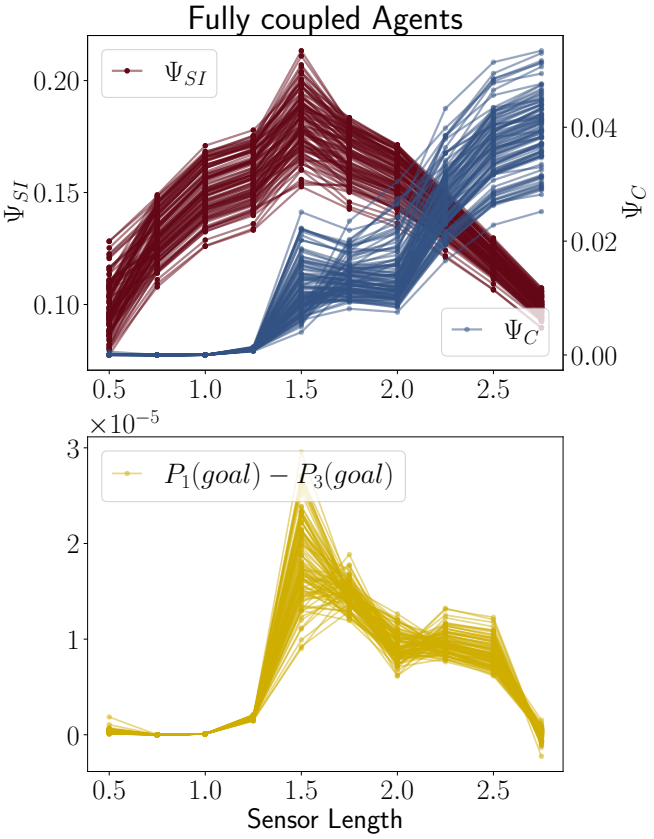


Figure 4.28: The measures Ψ_{SI} and Ψ_C in case of the fully coupled agents are displayed on the top and the bottom shows the difference in the probability of achieving the goal between the fully coupled and reactive control agents.

As hypothesized in the introduction in Section 4.1, information that is integrated in a meaningful manner in the controller should depend on information received from the sensors that at the same time makes an impact on the actuators.

The bottom graph in Figure 4.28 displays the difference between the probability of achieving the goal for the fully coupled agents, denoted by $P_1(goal)$, opposed to the reactive control agents, written as $P_3(goal)$. Comparing these results to Ψ_{SI} and Ψ_C above, reveals that the fully coupled agents perform much better in the cases, where Ψ_{SI} and Ψ_C are both high.

Therefore the importance of the information flow in the controller of an embodied agent depends additionally on the information flowing to and from the controller. Hence, only calculating the Integrated Information value Φ_T does not suffice to conclude how much of an impact the information integrated in the controller has.

We support this point further by examining the controller driven agents in more detail. In order to understand the information flow in the controller driven agents we consider the dynamics of the measures $\Psi_{AE}, \Psi_{Syn}, \Psi_C$ and Ψ_{SI} depicted in Figure 4.29. All of these measures are relatively high at a sensor length of 0.5 and have a maximum at 0.75. This results in a spike of the total information flow, Ψ_{TIF}, at 0.75, as visible in Figure 4.8.

The difference between the success of the fully coupled and controller driven agents, denoted by $P_2(goal)$, is depicted on the bottom of Figure 4.29. The black line marks the value 0. For a sensor length of 0.5, the fully coupled agents perform better than the controller driven ones, because the former are able to use reactive control. This allows a fast reaction to sensory input that is necessary in the case of short sensors.

However, for sensor with a length of 0.75 the controller driven agents achieve better results compared to the fully coupled agents. In order to understand the agents better we look at an exemplary distribution $P(A_{t+1}|C_t)$ that one of agents converged to, printed below in the Table 4.1. The first three columns indicate the states of C_t^1, C_t^2 and A_{t+1}^i and the last two columns are the rounded probabilities.

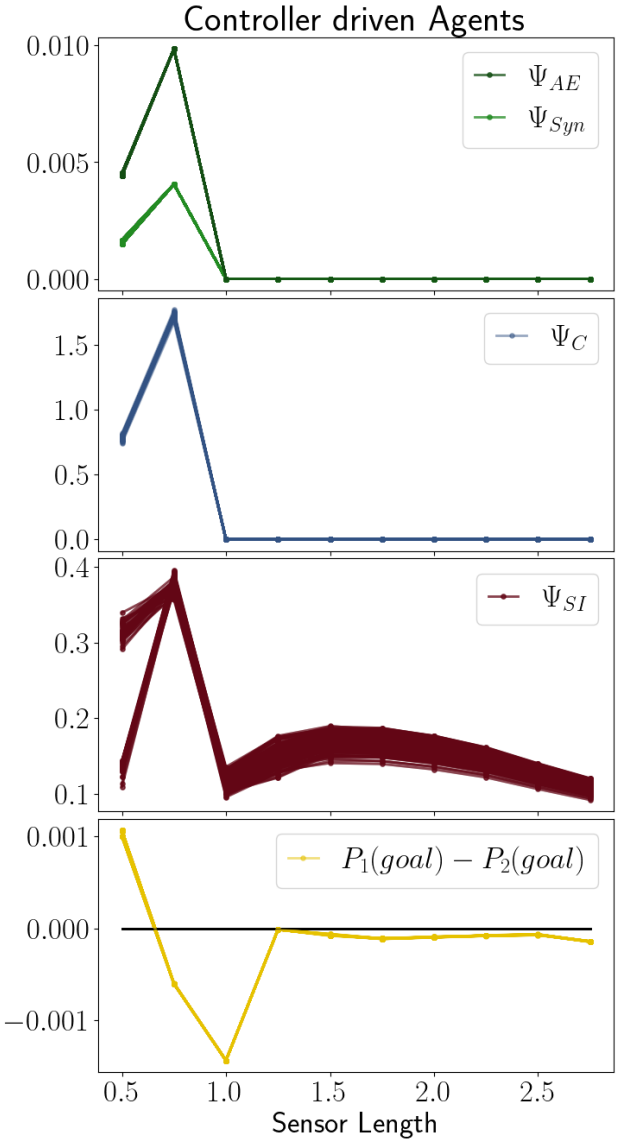


Figure 4.29: The measures Ψ_{AE}, Ψ_{Syn}, Ψ_C, Ψ_{SI} for the controller driven agents and the difference between the success of the fully coupled and controller driven agents.

c_t^1	c_t^2	a_{t+1}^i	$P(a_{t+1}^1\|c_t)$	$P(a_{t+1}^2\|c_t)$
0	0	0	≈ 1.39e-79	≈ 5.31e-80
0	0	1	≈ 1.00e+00	≈ 1.00e+00
0	1	0	≈ 1.00e+00	≈ 1.00e+00
0	1	1	≈ 1.95e-40	≈ 1.06e-40
1	0	0	≈ 1.00e+00	≈ 1.00+00
1	0	1	≈ 4.11e-11	≈ 2.48e-10
1	1	0	≈ 9.99e-01	≈ 9.99e-01
1	1	1	≈ 3.47e-09	≈ 2.93e-09

Table 4.1: One example of the distribution $P(A_{t+1}|C_t)$.

There we are able to observe that the agents move always slowly forward, except for the case where both controller states are 0, then the agents moves fast forward. If the agent would almost always move slowly forward, then the action would not depend on the controller at all. This is exactly what happens for sensors longer than 0.75. Note that in this case the agents converge to an almost deterministic distribution. Since the agent manifold, $\mathcal{M}_A$, consists of strictly positive distributions, however, one movement cannot be completely certain.

From a sensor length of 1 and up, the measures for action effect, synergistic information and control are very close to zero, with Ψ_{AE} between 10^{-6} and 10^{-10} and Ψ_C mostly between 10^{-4} to 10^{-16}. Therefore we come to the conclusion that the agents converged to an optimum in which they only move slowly forward so that the actuators do not depend on the sensory input and have no influence on the next sensory state.

Even though Φ_T has no impact on the actions of the agent, it still exhibits the decreasing behavior with increasing sensor length. Hence this supports once again that we should take the values of the sensory information and control into consideration when we judge the importance of the Integrated Information. In order to highlight this explicitly we next define the effective information integration.

Definition 20. (Effective Information Integration) An indicator for the impact of the Integrated Information on the behavior of an agent is defined as the product of the ground truth Integrated Information, the sensory information and control:

$$\Phi_{EII} = \Phi_T \cdot \Psi_{SI} \cdot \Psi_C.$$

It is called "effective information integration".

We choose to multiply these measures in order to assure that Φ_{EII} is zero if one or more of the measures Φ_T, Ψ_{SI} or Ψ_C are zero. Thereby we ensure that the effective information integration is only high when the integration of sensory data in the controller has an impact on the behavior of the agents.

The values of Φ_{EII} for the fully coupled and controller driven agents are depicted in Figure 4.30. The results in the case of the reactive control agents are zero, because the measure for control is always zero per definition. Here the controller integrates information, but it has no influence on the actuators and therefore the behavior of the agents.

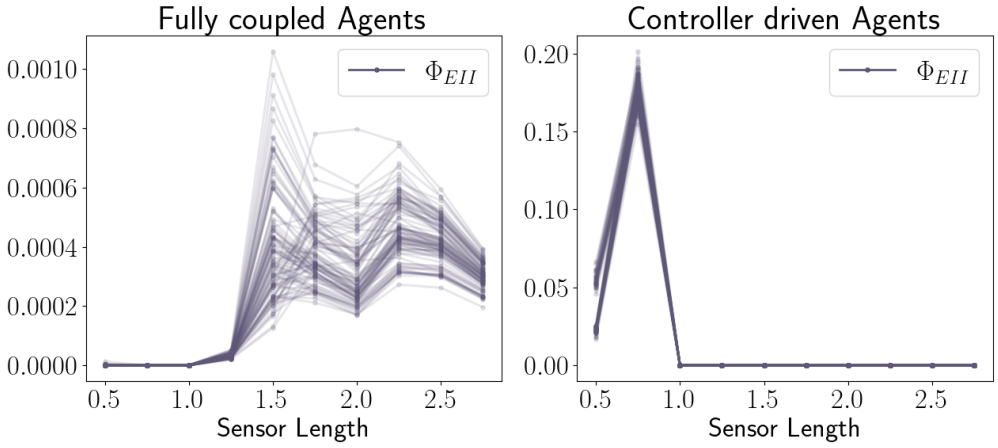


Figure 4.30: The results for the effective information integration in case of the fully coupled and controller driven agents.

The same holds true for the controller driven agents with sensors longer than 0.75, as discussed in the context of Figure 4.29.

In the case of the fully coupled agents, the results for the effective information integration look similar to the difference in the probabilities of achieving the goal between the fully coupled and reactive control agents, depicted on the bottom of Figure 4.28.

Therefore, the effective information integration describes the advantage that an agent has by being capable of integrating information in its controller.

4.7 Summary and Discussion of Chapter 4

In this chapter we analyze simulated agents in a racetrack. The goal of these agents is to maximize the probability of being alive after the next two movements. The optimization of the conditional distributions that control the agents is done by using the *em*-algorithm. This results in a very controlled setting, where the only variation between the agents comes from the random initial distributions that lead to different local optima. So, the outcomes for all considered 12 measures only vary slightly.

Hence, in this setting we are able to detect even minor differences between the three types of agents, namely the fully coupled, controller driven and reactive control agents. From the comparison among them we can conclude that the reactive control, Ψ_R, has a greater influence on the behavior of the agents than control, Ψ_C. It follows that the fully coupled and reactive control agents have a very similar information flow, since they only differ in the existence of a connection from the controller to the actuator nodes. The overall benefit of being able to integrate information is small, whereas the controller driven agents make no use of the controller at all for sensors longer than 1.

This points to a caveat of our setting. The architecture of the agents makes it necessary to consider two timesteps in order for the Integrated Information in the controller to have an influence on the actuator. Considering that the sensors can directly send their information to the actuators in only one timestep, utilizing reactive control, the connection through the controller might be too indirect. This might impact the possible influence of the controller on the actuators negatively.

Additionally, the agents have a very limited influence on their next sensory state over all. So their ability to affect whether they will reach the goal is reduced. The main reason for this comes from the decision to sample the empirical world model, $\tilde{P}(S_{t+1}|S_t, A_t)$, beforehand and then to apply the *em*-algorithm to the whole resulting distribution. Therefore the agents do not actually move inside the racetrack, but solely rely on the theoretical knowledge from the empirical world model in order to maximize the likelihood of success. It follows that the behavior of the agents has no impact on the world model in any way.

On the other hand, using this strict framework with little variations reveals clearly the significance of the change in the reach of the sensors. Varying the reach of the sensors has an immediate influence on the way the agents are able to interact with their environment, quantified by a change in the measured Morphological Computation. This allows us to observe the antagonistic relationship between the measure for Integrated Information, Φ_T, and the Morphological Computation, Ψ_{MC}. Hence, the information flow on the inside of the agent has an inverted dynamic compared to the information flow through the world, on the outside of the agent. In short, the more the agent interacts with its environment, the less information is integrated.

This relationship leads to the following problem. It insinuates that embodied intelligence could be correlated with reduced conscious experience. Hence, this leads to the question why agents with a well-adapted morphology would need Integrated Information at all. Would it not be possible to build agents that are so well adapted to their environment that any

Integrated Information is irrelevant?

Furthermore, our experiments demonstrate that even in the cases where the Integrated Information is high this alone is not sufficient to conclude that the behavior of the agent is influenced by it. We observe that an agent can have a high Integrated Information value that has no influence on its actuator states. Hence, we analyze additionally the information flow to and from the controller, measured by Ψ_{SI} and Ψ_C. In order to ensure that we consider only the Integrated Information that processes the information from the sensors and then influences the actions of the agent accordingly, we additionally introduce the effective information integration, Φ_{EII}.

The effective information integration makes the posed problem even more evident. It seems that in most cases the agents perform well without any interference from a controller. Hence, why would we consider a controller, in general, or Integrated Information, in particular, to be necessary?

It might be that the agents themselves and problems they are faced with here are too simplistic. Posing a more complicated exercise could lead to agents that have to make use of their controller in order to reach their goal.

In the next chapter we build on the results presented here and propose an alternative solution, given by the challenge of learning. The sailor from the introductory example in Section 4.1 might not consider navigating a difficult mental task, because he is well trained in doing so. Learning to navigate, on the other hand, most likely required far more conscious considerations.

Hence, in the following chapter we make adjustments to the setting of the experiment that alleviate the discussed caveats and then we analyze the behavior of the agents during learning. There the agents have to learn the optimal behavior as well as to understand their individual environments directly inside the racetrack.

5 The Information Flow in a Learning Agent

The previous chapter discusses experiments in which we analyze the information flow in agents that perform goal directed actions. We continue the analysis in this chapter by modifying the experiment to closely observe the learning process of agents that act in their environment. Here, the agents have to simultaneously learn an optimal behavior and a prediction of the next sensory state. The latter is called an internal world model.

We address the concept of a world model in more detail in Section 5.1.1 after the introduction of our motivation. The modified experimental setup is presented in Section 5.2. In order to include the learning of the world model we adjust the *em*-algorithm, as explained in Section 5.4. Since the agents have an internal world model, we define four additional measures on this model in Section 5.5. Finally, we discuss the results of the experiments in Section 5.6. Those are partly included in the preprint [Langer24].

5.1 Introduction

In the previous chapter we discuss an antagonistic relationship between Morphological Computation and Integrated Information. The experiments show that Morphological Computation, the reduction of computational cost for the controller resulting from body-environment interactions, can be sufficient for an agent to reach simple goals. This depends on how well adapted the body is to the environment and task. Hence, the question arises whether a controller is only necessary for more complicated tasks or, more precisely, does a sufficiently well-designed body alleviate the necessity for Integrated Information and therefore for conscious experience? Here we address one possible answer given by the challenge of learning to act in an environment.

For every embodied agent learning to perform new tasks and navigating in their environments is demanding. An important aspect of this task is that they need to be able to predict what is happening next and especially what the outcomes of their own actions would be. In order to do that the agents need a world model, which we discuss in more detail in the next section. We illustrate the intuition behind the assumption that Integrated Information might be necessary for learning with following example.

Consider a child who learns how to ride a bike. Nearly every task the child has learned up to this point, e.g. walking, speaking or drawing, becomes harder when one tries to do it fast. Therefore, the child most likely expects that moving slowly would be the best approach to this problem. According to its understanding of the world, its internal world model, riding a bike slowly is easier than doing it fast. Unfortunately this is not the case here. Speed stabilizes a bike and is therefore beneficial for cycling. The child is working with an inaccurate world model, which complicates the problem further. Hence, before the child can accomplish to ride a bike it has to try, fail, observe and understand that faster sometimes can mean easier. It has to update its world model in order to learn and to be able to use the world in an optimal way.

In this chapter we closely examine the process of updating the world model and analyze learning agents with an optimal versus a sub-optimal world model. Thereby we observe how the information is integrated in the controller during this learning process. We manipulate the accuracy of the world model, which leads us to our first result:

1. An agent that understands its environment, meaning it has an accurate world model, exhibits a higher Morphological Computation and lower controller complexity compared to agents with an inaccurate world model. The better an agent understands its environment, the more it can exploit the interactions between body and environment and the less controller complexity is needed.

Following the results from the previous chapter, we quantify the impact of the Integrated Information on the behavior of the agent by the effective information integration. Additionally, some agents have to form an internal world model in order to predict their next sensory state. The complexity of the internal world model is measured by the synergistic prediction and this adds to the complexity of the controller. Hence, the controller complexity now consists of two parts, the Integrated Information and the synergistic prediction.

In the experiments we observe that agents that succeed at learning the task have first a high controller complexity and then this value decreases. We hypothesize that this is because the agents first have to learn the correct world model before they are able to optimally utilize the interaction of their bodies with the environment, measured by Morphological Computation, which in turn leads to a lower Integrated Information. This is supported by the result that unsuccessful agents have a constantly high Integrated Information and a lower Morphological Computation, compared to the successful agents.

Additionally, we examine agents that are not able to integrate information in the controller and call them "split" agents. Those agents perform significantly worse. By calculating four different measures for the internal world model we discover that the few successful split agents combine their different information sources directly in the world model. Note that this implies that there is an information integration in the world model, which leads us to the following conclusion:

2. In order to learn a successful behavior the agents have to combine information from different sources. This leads to an increased controller complexity either directly between the controller nodes, in form of Integrated Information, or in the prediction process given by the internal world model.

Before we present the details of our experiments we first review the concepts of "world model" and "prediction".

5.1.1 The World Model and Prediction

In 1970 Conant and Ashby discussed in [Conant70] the necessity of a model for a regulator of a system and conclude with the following resolution.

> "There can no longer be question about *whether* the brain models its environment: it must."

Later, in 1976, Francis and Wonham formalized in [Francis76] their "Internal Model Principle of Control Theory". This states that an internal model is necessary for a control architecture in a regulator problem. They show, under some regularity assumptions, that in the case of linear multivariate systems an internal world model is required for a controller to work.

In general, an internal world model mimics the dynamics of a more complex, external process. Empirical support for the existence of an internal model in humans can be found in many studies, for example in [Wolpert95b]. The concept of an internal world model has been further developed and formalized in the context of control theory, see for example [Wolpert95a; Jordan96; Braun09].

A well known distinction can be made between two types of internal models, namely forward and inverse models, [Jordan96; Cooper10]. A forward model takes the motor actions, or an efferent copy of the motor actions, as input and predicts the next sensory states, e.g. the position of an agent. The difference between the actual outcome and the prediction can then be used to improve the motor commands.

As the name suggests, an inverse model describes the dynamic of the action taken with the next sensory state as the input. This can be gained by inverting the forward model or by learning two separate processes. An inverse model can be used for control by using the required outcome as an input variable. The inverse model then produces the action that could lead to the desired state. This method is called "predictive processing", as described in [Jordan96].

Here, we call the forward model, the mechanism that generates a prediction, the "internal world model", which we define in detail in Section 5.4. In the previous section the agents have access to the sampled empirical world model, which is an accurate description of the dynamics of the agents environments. The internal world model then mimics the dynamics of the more complex, empirical world model. The empirical world model was termed world model in [Ghazi-Zahedi10] and [Montúfar15]. The internal prediction was also named "world model" in [Ay13b; Ay14] and optimal world model with respect to its prediction qualities in [Ay13a]. To prevent confusion, we make the explicit distinction between empirical and internal world model.

5.2 Setting of the Experiment

Here we describe the setting of our experiments, which is similar to the one introduced in Section 4.2. We analyze the information flow of simplistic, 2-dimensional agents in a racetrack. Hereby, we use the same racetrack as in the previous chapter, but we introduce changes affecting the dynamics, body, architecture and goal of the agents to address the shortcomings of our previous experiments.

The objective of the agents is to move inside the racetrack without touching the walls. Whenever the body of an agent touches a wall the agent gets stuck. This means that it can only turn on the spot but will not move away unless both sensors do not detect a wall. Figure 5.1 on the top depicts a section of the racetrack with an agent that got stuck twice. There every 5th step is printed in darker colors to increase the visibility. At an x-axis value of 5 the agents body touches the wall and turns red, then the agent has to turn on the spot, visible by the different red tails, before it can move away. A video of an agent moving inside a racetrack can also be found at [Langer22].

The movement of the agents is almost the same as described in Section 4.2, namely fast forward, slow forward, left and right, but now the agents are able to turn with approx. 14° instead of 10°. This reduces the number of steps that the agents are stuck at a wall while turning.

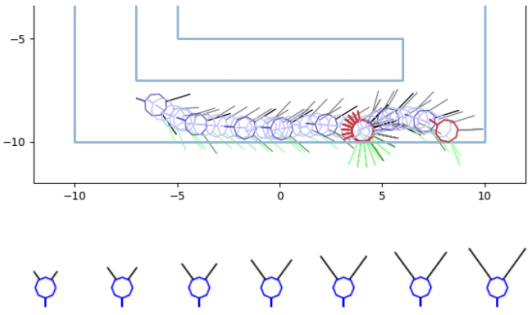


Figure 5.1: An agent moving inside the racetrack on the top and the possible sensor length from 0.5 on the bottom left to 2 on the bottom right.

In order to prevent an agent from going around in circles without moving through the racetrack we enlarge the agent's body from 0.3 to 0.55.

In addition, we also vary the reach of the sensors from 0.5 to 2 in steps of 0.25, as depicted on the bottom of Figure 5.1. As discussed previously, this directly influences the amount of information an agent receives about the world and hence it impacts the quality of the interaction of the agent with its environment. This is measured by Morphological Computation, as discussed in Section 4.2 and 4.1.1.

In the next section we introduce the architecture of the agents in detail.

5.3 The Agents and the World Model

An agent in our experiment is modeled by a discrete multivariate, time-homogeneous Markov process

$$(X_t)_{t\in\mathbb{N}} = (S_t, A_t, C_t)$$

with the state space $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{C}$. Here the variable S_t describes the two binary sensors that detect a wall and additionally a binary variable that encodes whether the body of the agent is touching a wall or not. The node A_t includes the two binary actuator and C_t the two binary controller nodes. Additionally, we introduce another variable here, called S'_t, that describes the internal prediction of the next sensor state and hence consists of three binary variables.

The elements of the agents are connected according to the graph in Figure 5.2. This graph then leads to a distribution that factorizes in the following way

$$P(x_t, x_{t+1}, s'_{t+1}) = P(s_t, a_t, c_t)P(s'_{t+1}|a_t, c_t)P(s_{t+1}|s_t, a_t) \cdot \prod_{j=1}^{2} P(c^j_{t+1}|c_t, s_{t+1}) \prod_{i=1}^{2} P(a^i_{t+1}|s_{t+1}, c_{t+1}),$$

for all $(x_t, x_{t+1}, s'_{t+1}) \in \mathcal{X} \times \mathcal{X} \times \mathcal{S}$. Here we only depict one node for each S, S', A and C in the figures in order to increase clarity.

Note that we not only introduce a new variable, but also change she structure of the process. The reason for this is that with the previous couplings the connections among the controller nodes have a delayed influence on the actuator nodes. Now the actuator depends on the sensor and controller values in the same point in time, the influences are instantaneous. This increases the importance of the controller.

Since S'_t is an internal prediction of S_t, it is made of the same substrate, hence the state space of S'_t is also $\mathcal{S}$. The difference between S_t and S'_t lies solely in the mechanism with which they are generated. The node S_t is influenced by the information from S_{t-1} and A_{t-1}. These are indirect influences, since the information flows through the environment, described by the empirical world model.

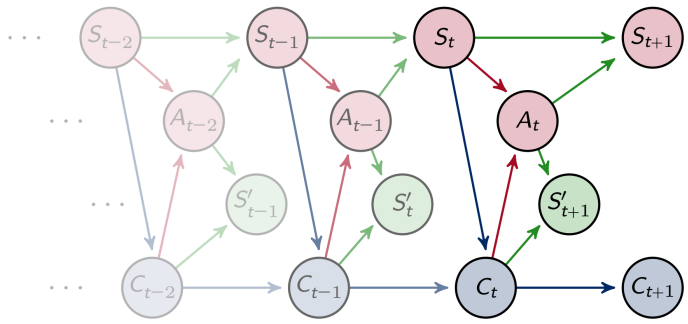


Figure 5.2: The sensorimotor loop of the learning agents.

The role of the environment is discussed further in Section 4.2.2. In that section we sample the empirical world model, $P(S_{t+1}|S_t, A_t)$, for every sensor length. Here, each agent samples their own empirical world model depending on the path they take inside the racetrack, as described in more detail in [Ghazi-Zahedi10]. In addition we sample the distribution for S_t, A_t and C_t and we denote the sampled distributions by $\tilde{P}(S_{t+1}|S_t, A_t)$ and $\tilde{P}(S_t, A_t, C_t)$.

Previously, the agents had direct access to the empirical world model and therefore no need for an internal world model. We also consider a set of agents here that do not need to form an internal world model, $P(S'_{t+1}|C_t, A_t)$. These agents sample their environment, meaning that they instantaneously and accurately include their experiences in their world

model. We refer to these agents as "ideal agents", since agents normally do not have direct access and a perfect understanding of their environment.

Furthermore, we divide the agents that need to learn an internal world model into two classes. The first type of agent is fully connected, meaning that both controller nodes depend on both controller nodes in the previous point in time. The second one is not able to generate Integrated Information in its controller. This is done by ensuring that the controller node C^j_{t+1} only receives information from C^j_t and not from $C^{j'}_t$, $j, j' \in \{1, 2\}$ with $j' \neq j$. We refer to them as fully connected and "split" agents.

5.4 Learning

Perhaps the most important difference between these experiments and the ones discussed in the previous chapter is that here the agent learns while it is inside the racetrack. Thereby its actions directly influence the way the agent perceives its environment.

At each step t, the realized states s_{t-1}, a_{t-1} and c_{t-1} are known. Hence, instead of considering all the different possibilities the agent can use these certainties. To that end we need the following definitions. Let $P_{a_t}(C_{t+1}|S_{t+1})$ be the probability distribution of C_{t+1} conditioned on S_{t+1} and a fixed state a_t:

$$P_{a_t}(c_{t+1}|s_{t+1}) := P(c_{t+1}|s_{t+1}, a_t),$$

for $s_{t+1}, c_{t+1} \in \mathcal{S} \times \mathcal{C}$.

Now we consider an internal, agent-centric perspective. For the ideal agents with access to the empirical world model the process is as depicted in Figure 5.3 on the top. Here the prediction of the next sensory state is denoted by $\hat{S}$. Note that the s_t is the actual realized last sensor state and not an internal prediction. The ideal agents optimize the distributions $P_{c_t}(C_{t+1}|\hat{S}_{t+1})$ and $P(A_{t+1}|\hat{S}_{t+1}, C_{t+1})$ using the *em*-algorithm, as discussed in Section 4.3.

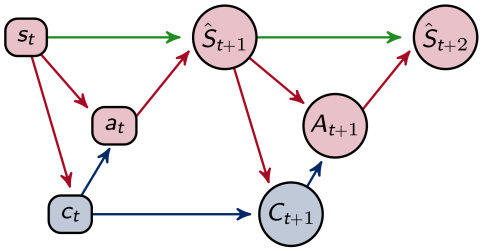


Figure 5.3: The sensorimotor loop from the perspective of the ideal agents.

The internal perspective of the agents with an internal world model is given in Figure 5.3 on the bottom. Considering this process the agent can optimize all the conditional distributions involved, namely $P_{c_t}(C_{t+1}|S'_{t+1})$, $P(A_{t+1}|S'_{t+1}, C_{t+1})$ and $P(S'_{t+2}|A_{t+1}, C_{t+1})$. Additionally, $P_{s_t,a_t}(S'_{t+1})$ can be gained from $P(S'_{t+2}|A_{t+1}, C_{t+1})$ because the process is time-homogeneous.

It is important to note that if the internal world model is not close to the empirical world model, then this process does not reflect the actual dynamics of the agent's movements. In that case, it cannot find an optimal solution for its actuator states, because the optimization of the behavior uses faulty assumptions leading to a failure of the agent. In the example in the introduction this would be the child trying to learn to ride a bike while going as slow as possible. Hence, both of the world models should result in similar predictions. Therefore, we include the learning of the internal world model in the learning process. We do so by modifying the *em*-algorithm that we applied in the previous experiments, described in Section 2.5.1 and Section 3.4.1. Next, we describe these modifications of the *em*-algorithm.

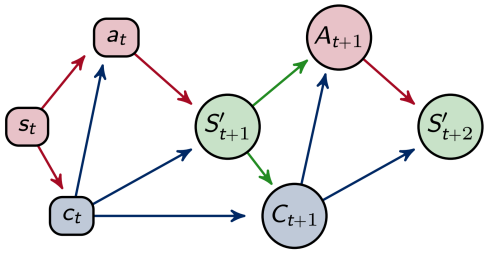


Figure 5.4: The sensorimotor loop from the perspective of the agents with an internal world model.

If the agents have to learn an internal world model, then they have two separate goals. On the one hand, they want to optimize the distributions $P_{c_t}(C_{t+1}|S'_{t+1})$ and $P(A_{t+1}|S'_{t+1}, C_{t+1})$ such that the probability of touching the wall after the next movement is as low as possible. On the other hand, they need to keep the prediction generated by the internal world model $P(S'_{t+1}|C_t, A_t)$ close to the result from the empirical world model $P(S_{t+1}|S_t, A_t)$. The two types of world models are highlighted in gold in Figure 5.5.

So we alternate between optimizing $P_{c_t}(C_{t+1}|S'_{t+1})$ and $P(A_{t+1}|S'_{t+1}, C_{t+1})$ with respect to the goal on one hand and with respect to the difference between $P(S'_{t+1}|C_t, A_t)$ and $P(S_{t+1}|S_t, A_t)$ on the other hand. Details of this optimization are discussed in the next section.

Note that the controller has only two binary variables, whereas the sensor consists of 3 binary variables. Therefore, merely copying the information from the sensor nodes is not a viable strategy for the agents. This is a natural assumption, because humans or animals are also not able to consciously perceive every detail from the environment that their sensors are able to pick up. Hence, one has to learn to distinguish between important and irrelevant information.

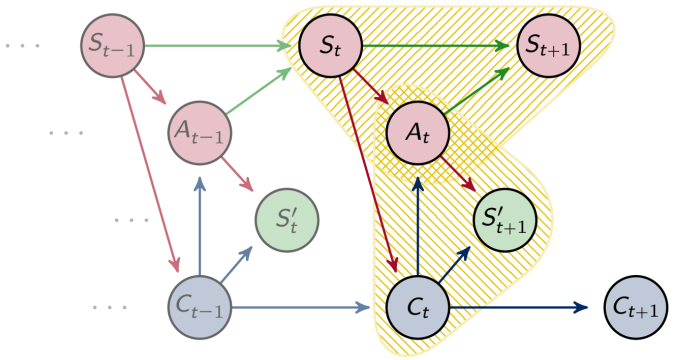


Figure 5.5: The sensorimotor loop with a highlighted empirical and internal world model.

Furthermore, we add Gaussian noise to the distribution $P(A_{t+1}|S'_{t+1}, C_{t+1})$, because if the *em*-algorithm reaches a point where for some action $P(a_{t+1}|s'_{t+1}, c_{t+1}) = 0$ holds, then it otherwise can not gain a positive value again.

5.4.1 Training the Behavior and the internal World Model simultaneously

The learning algorithm, applied in our experiments to agents with an internal world model, works by adapting the *em*-algorithm to incorporate two different goals. In this case the agents need to optimize their behavior with respect to the goal of not touching a wall as well as their internal world model. Hence, now we define four different types of sets between which we iterate.

The first of these sets is defined for optimizing with respect to reaching the goal of not being stuck at a wall. Let S'^3 be the variable indicating whether the agent is touching a wall. Then $s^3 = 1$ indicates that the agent is not touching a wall. Now, we define the random variable $X'_{t+1} = (S'_{t+1}, A_{t+1}, C_{t+1})$. The goal manifold then consists of all those probability distributions for which it is certain that the agent will not touch a wall at time point $t + 2$:

$$\mathcal{M}_G^P(x_t) = \left\{P \in \mathcal{P}(\mathcal{X} \times \mathcal{S}) | P(s^3_{t+2} = 1) = 1\right\}.$$

This is very similar to the set $\mathcal{M}_G$ in Section 4.3.1. Here, the upper index P stands for "policy" since this is the goal manifold with which the behavioral rules are optimized.

The second set follows the same reasoning as the agent manifold in Section 4.3.1. It consists of all the distributions that factor according to the agents. This means that the set

$\mathcal{M}_A^P$ contains all of the possible agents given the current world model:

$$\mathcal{M}_A^P(x_t, \overline{P}) = \Bigg\{ Q \in \mathcal{P}^\circ(\mathcal{X} \times \mathcal{S}) | Q(x_{t+1}, s_{t+2}) = \overline{P}_{a_t, c_t}(s_{t+1}) \prod_j Q_{c_t}(c_{t+1}^j | s_{t+1}) \prod_i Q(a_{t+1}^i | s_{t+1}, c_{t+1}) \overline{P}(s_{t+2} | a_{t+1}, c_{t+1}),\ \forall (x_{t+1}, s_{t+2}) \in \mathcal{X} \times \mathcal{S} \Bigg\},$$

where $\mathcal{P}^\circ$ is the interior of $\mathcal{P}$ and $\bar{P}$ indicates that this distribution is fixed.

In Section 4.3.1 we iteratively project between these two sets in order to find the distribution in $\mathcal{M}_A^P$ that is closest to $\mathcal{M}_G^P$. This would be the distribution that describes a valid agent and has a high likelihood of achieving the goal. We apply this method to the ideal agents in which case $\bar{P}$ in $\mathcal{M}_A^P(x_t, \bar{P})$ is given by the empirical world model.

Here, we want to adapt this approach in order to simultaneously learn the internal world model. The distribution $P(S'_{t+1} | A_t, C_t)$ predicts the next sensory input and reflects therefore the agent's understanding of its environment. Hence we want to optimize our world model such that

$$P(S'_{t+2} | S_{t+1}, A_{t+1}) = \tilde{P}(S_{t+2} | S_{t+1}, A_{t+1}),$$

where $\tilde{P}$ is the sampled, empirical world model.

Note that we require the goal to be a restriction on a joint distribution, not a conditional, in order to define the e-projection to this goal manifold. Hence, the actual optimization considers the following equality

$$\overline{P}(s_{t+1}, a_{t+1}) P(s_{t+2} | s_{t+1}, a_{t+1}) = \overline{P}(s_{t+1}, a_{t+1}) \tilde{P}(s_{t+2} | s_{t+1}, a_{t+1}),$$

for all $(s_{t+1}, s_{t+2}, a_{t+1}) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$. The joint distribution $\overline{P}(S_{t+1}, A_{t+1})$ is fixed to the joint distribution from the previous step in the algorithm.

Then the conditional distribution $P(S'_{t+2} | S_{t+1}, A_{t+1})$ can be calculated using the conditional distributions known to the agent as follows

$$P(s_{t+2} | s_{t+1}, a_{t+1}) = \frac{\sum_{c_{t+1}} P_{a_t, c_t}(s_{t+1}) P_{c_t}(c_{t+1} | s_{t+1}) P(a_{t+1} | s_{t+1}, c_{t+1}) P(s_{t+2} | c_{t+1}, a_{t+1})}{\sum_{c_{t+1}} P_{a_t, c_t}(s_{t+1}) P_{c_t}(c_{t+1} | s_{t+1}) P(a_{t+1} | s_{t+1}, c_{t+1})},$$

for all $(s_{t+1}, s_{t+2}, a_{t+1}) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$. This allows us to define the third set, called the world goal manifold, which is again a linear family similar to the goal manifold in Section 4.3.1:

$$\mathcal{M}_G^W(x_t, \overline{P}) = \Big\{ P \in \mathcal{P}(\mathcal{X} \times \mathcal{S}) | P(s_{t+2}, s_{t+1}, a_{t+1}) = \overline{P}(s_{t+1}, a_{t+1}) \tilde{P}(s_{t+2} | s_{t+1}, a_{t+1}),\ \forall (s_{t+1}, s_{t+2}, a_{t+1}) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A} \Big\}.$$

The last set is similar to the agent manifold above and we call this the world agent manifold because it consists of all the possible agents given the conditional distributions on C_{t+1} and A_{t+1}

$$\mathcal{M}_A^W(x_t, \overline{P}) = \Bigg\{ Q \in \mathcal{P}^\circ(\mathcal{X} \times \mathcal{S}) | Q(x_{t+1}, s_{t+2}) = Q_{a_t, c_t}(s_{t+1}) \prod_j \overline{P}_{c_t}(c_{t+1}^j | s_{t+1}) \prod_i \overline{P}(a_{t+1}^i | s_{t+1}, c_{t+1}) Q(s_{t+2} | a_{t+1}, c_{t+1}),\ \forall (x_{t+1}, s_{t+2}) \in \mathcal{X} \times \mathcal{S} \Bigg\}.$$

Since both agent manifolds, $\mathcal{M}_A^P$ and $\mathcal{M}_A^W$, vary only through the fixed parts of the distributions, we can define a full agent manifold by

$$\mathcal{M}_A(x_t) = \Bigg\{ Q \in \mathcal{P}^\circ(\mathcal{X}\times\mathcal{S}) | Q(x_{t+1}, s_{t+2}) = Q_{a_t,c_t}(s_{t+1}) \prod_j Q_{c_t}(c_{t+1}^j | s_{t+1}) \prod_i Q(a_{t+1}^i | s_{t+1}, c_{t+1}) Q(s_{t+2} | a_{t+1}, c_{t+1}),\ \forall (x_{t+1}, s_{t+2}) \in \mathcal{X}\times\mathcal{S} \Bigg\}.$$

Then the inclusions $\mathcal{M}_A^W \subset \mathcal{M}_A$ and $\mathcal{M}_A^G \subset \mathcal{M}_A$ hold. Similarly, we define a full world goal manifold

$$\mathcal{M}_G^W(x_t) = \Bigg\{ P \in \mathcal{P}(\mathcal{X}\times\mathcal{S}) | \exists R \in \mathcal{P}(\mathcal{X}) : P(s_{t+2}, s_{t+1}, a_{t+1}) = R(s_{t+1}, a_{t+1}) \tilde{P}(s_{t+2} | s_{t+1}, a_{t+1}),\ \forall (s_{t+1}, s_{t+2}, a_{t+1}) \in \mathcal{S}\times\mathcal{S}\times\mathcal{A} \Bigg\}.$$

This set includes all world goal manifolds.

Although we have two linear families expressing our goals, namely $\mathcal{M}_G^P(x_t)$ and $\mathcal{M}_G^W(x_t)$, it is in general not possible to apply the iterative scaling algorithm to compute an e-projection to their intersection. In most applications, including the experiments described here, there exists no empirical world model that guarantees success. Hence, the intersection of $\mathcal{M}_G^P(x_t)$ and $\mathcal{M}_G^W(x_t)$ is empty and therefore we need to modify the *em*-algorithm in a different manner.

Now we define the algorithm depicted in Figure 5.6. The first step of the modified *em*-algorithm is to project to $\mathcal{M}_G^P(x_t)$ via an e-projection

$$P^0 = \operatorname*{arg\,inf}_{P \in \mathcal{M}_G^P(x_t)} D_{\mathcal{X}\times\mathcal{S}}(P \parallel Q^0).$$

Then we project the resulting distribution with an m-projection to $\mathcal{M}_A^G(x_t, Q^0)$

$$Q^1 = \operatorname*{arg\,inf}_{Q \in \mathcal{M}_A^P(x_t, Q^0)} D_{\mathcal{X}\times\mathcal{S}}(P^0 \parallel Q).$$

Up to this point, this is the standard *em*-algorithm, as it was used in Section 4.3.1.

Now instead of projecting to $\mathcal{M}_G^P(x_t)$ again, we update the internal world model by projecting to $\mathcal{M}_G^W(x_t, Q^1)$ with an e-projection:

$$P^1 = \operatorname*{arg\,inf}_{P \in \mathcal{M}_G^W(x_t, Q^1)} D_{\mathcal{X}\times\mathcal{S}}(P \parallel Q^1).$$

Afterwards we project this P^1 with an m-projection to $\mathcal{M}_A^W(s_t, a_t, c_t, Q^1)$:

$$Q^2 = \operatorname*{arg\,inf}_{Q \in \mathcal{M}_A^W(x_t, Q^1)} D_{\mathcal{X}\times\mathcal{S}}(P^1 \parallel Q).$$

Now we projected to each type of set, namely $\mathcal{M}_G^P$, $\mathcal{M}_A^P$, $\mathcal{M}_W^G$ and $\mathcal{M}_A^W$, exactly once. Then we start the whole process over by projecting to $\mathcal{M}_G^P(x_t)$ again, as depicted in the Sketch in Figure 5.6. Thereby we cycle through the projections and improve the internal world model and the agent's behavior simultaneously.

Note that this modified algorithm is not guaranteed to converge. Since we are interested in agents that learn while performing a task, we execute only a few optimization steps, as described above, after each step of the agent. More precisely, we project to each of the four sets exactly five times after every movement of the agent. Therefore, the agents adapt their behavior and internal world model after each step slightly and a convergence is not needed in our scenario. It is important to to stop this algorithm after a projection to $\mathcal{M}_A(x_t)$ since we want to find conditional probability distributions that describe a valid agent.

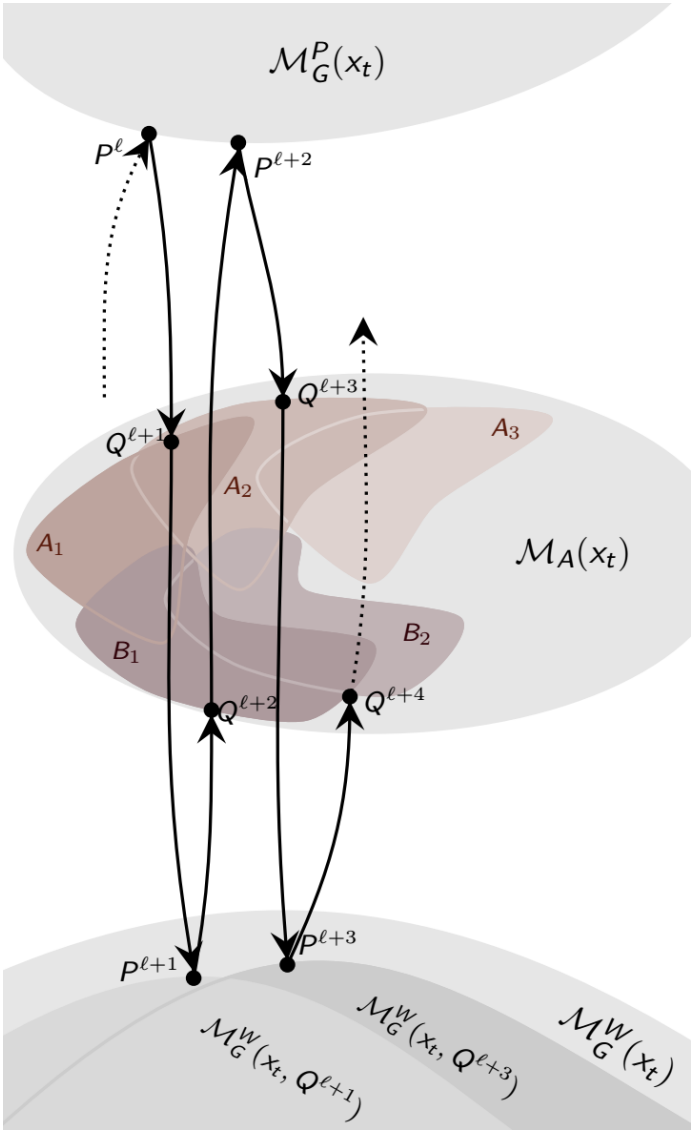


Figure 5.6: Sketch of the modified *em*-algorithm for optimizing the behavior and the internal world model simultaneously. Here $A_1 = \mathcal{M}_A^P(x_t, Q^l)$, $A_2 = \mathcal{M}_A^P(x_t, Q^{l+2})$, $A_3 = \mathcal{M}_A^P(x_t, Q^{l+4})$, $B_1 = \mathcal{M}_A^W(x_t, Q^{l+1})$ and $B_2 = \mathcal{M}_A^W(x_t, Q^{l+3})$.

It remains to define the different projections. These *e*- and *m*-projections here are very similar to the *e*- and *m*- projections proven in Section 4.3.1, respectively. Therefore, we give the results of the minimizations here directly and omit the proofs.

The e-projection $P_P^\star$ of Q to $\mathcal{M}_G^P$ results in

$$\underset{P\in\mathcal{M}_G^P(x_t)}{\arg\inf} D_{\mathcal{X}\times\mathcal{S}}(P \parallel Q^0) = P_P^\star(x_{t+1}, s_{t+2})$$

$$P_P^\star(x_{t+1}, s_{t+2}) = Q(x_{t+1}, s_{t+2})\frac{P(s_{t+2}^3)}{Q(s_{t+2}^3)},$$

for all $(x_{t+1}, s_{t+2}) \in \mathcal{X}\times\mathcal{S}$. Similarly, the e- projection $P_W^\star$ of Q to the world goal manifold, $\mathcal{M}_G^W$, is given by

$$\underset{P\in\mathcal{M}_G^W(x_t,Q^1)}{\arg\inf} D_{\mathcal{X}\times\mathcal{S}}(P \parallel Q) = P_W^\star(x_{t+1}, s_{t+2})$$

$$P_W^\star(x_{t+1}, s_{t+2}) = Q(x_{t+1}, s_{t+2})\frac{P(s_{t+2}, s_{t+1}, a_{t+1})}{Q(s_{t+2}, s_{t+1}, a_{t+1})},$$

for all $(x_{t+1}, s_{t+2}) \in \mathcal{X}\times\mathcal{S}$.

Now we define the m-projections. Minimizing the KL-divergence with respect to the second argument between P and the elements in $\mathcal{M}_A^P$ leads to the distribution $Q_P^\star$

$$\underset{Q\in\mathcal{M}_A^P(x_t,\bar{P})}{\arg\inf} D_{\mathcal{X}\times\mathcal{S}}(P \parallel Q) = Q_P^\star$$

defined by

$$Q_P^\star(x_{t+1}, s_{t+2}) = \overline{P}_{a_t,c_t}(s_{t+1})\prod_j P_{c_t}(c_{t+1}^j|s_{t+1})\prod_i P(a_{t+1}^i|s_{t+1}, c_{t+1})\overline{P}(s_{t+2}|a_{t+1}, c_{t+1}),$$

for all $(x_{t+1}, s_{t+2}) \in \mathcal{X}\times\mathcal{S}$.

Lastly, the m-projection $Q_W^\star$ of P to $\mathcal{M}_A^W$ results in

$$\underset{Q\in\mathcal{M}_A^W(x_t,\bar{P})}{\arg\inf} D_{\mathcal{X}\times\mathcal{S}}(P \parallel Q) = Q_W^\star$$

with

$$Q_W^\star(x_{t+1}, s_{t+2}) = P_{a_t,c_t}(s_{t+1})\prod_j \overline{P}_{c_t}(c_{t+1}^j|s_{t+1})\prod_i \overline{P}(a_{t+1}^i|s_{t+1}, c_{t+1})P(s_{t+2}|a_{t+1}, c_{t+1}),$$

for all $(x_{t+1}, s_{t+2}) \in \mathcal{X}\times\mathcal{S}$. An implementation of this algorithm can be found in [Langer22].

In conclusion, this modified version of the *em*-algorithm optimizes the goal of not touching the walls and the understanding of the world iteratively. After the optimization in each step we then apply different information theoretic measures to the resulting distribution. We define the applied measures in the next section.

5.5 Measures of the Information Flow in the internal World Model

In this section we introduce four information theoretic measures for analyzing the information flow in the internal world model. Additionally, we include six of the measures that we discussed in the previous chapter in this setting.

We measure the importance of an information flow by calculating the difference between the actual distribution and the closest distribution without the information flow in question, as explained in Definition 19.

The first measure for the prediction process, defined by the internal world model $P(S'_{t+1}|A_t, S_t)$, is called "full prediction" Ψ_{FP}. This quantifies the information flow from C_t and A_t to S'_{t+1} and the measure results in the mutual information between S'_{t+1} and (A_t, C_t)

$$\Psi_{FP} = I(S'_{t+1}; A_t, C_t).$$

The graphical representation corresponding to the split system of the full prediction is depicted in Figure 5.7.

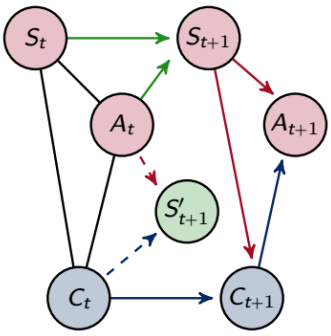


Figure 5.7: Graphical representation of the split system of Ψ_{FP}.

Analogously to the measure quantifying the total information flow Ψ_{TIF}, discussed in Section 4.5.1, the full prediction is an upper bound for the three other measures that we define next.

We start with two measures that each remove exactly one of the connections from C_t or A_t to S'_{t+1}. First, we discuss the "actuator prediction". This calculates the influence the actuator has on the prediction of the next sensory state.

The graphical representation of the corresponding split system is depicted in Figure 5.8 and the measure can be calculated as follows

$$\Psi_{AP} = I(S'_{t+1}; A_t|C_t).$$

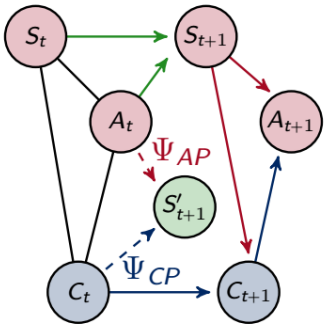


Figure 5.8: Graphical representation of the split systems of Ψ_{AP} and Ψ_{CP}.

Equivalently, we define the "controller prediction" by removing the connection between the prediction of the next sensory state and the controller.

$$\Psi_{CP} = I(S'_{t+1}; C_t|A_t).$$

This describes how dependent the prediction of the next sensory state is on the controller nodes, given the actuator states. The graphical representation of the split system is also shown in Figure 5.8.

The last of the measures for the internal world model quantifies how important the interplay between the influences of A_t and C_t on S'_{t+1} are. Hence, this measures a type of information integration of the information flows from A_t and C_t to S'_{t+1}. Therefore, we consider this measure as quantifying the complexity of the internal world model.

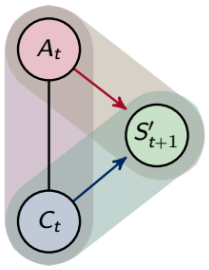


Figure 5.9: Sketch of the hierarchical model corresponding to the split system in case of Ψ_{SynP}.

This measure has no closed form solution and is conceptually similar to the synergistic information, Ψ_{Syn}, defined in Section 4.5.3. Here we define a split system where we only allow the two-way interactions among A_t, C_t and S'_{t+1}, but no combined influence from (A_t, C_t) on S'_{t+1}. Hence, we call this measures "synergistic prediction", Ψ_{SynP}.

The split system results in a hierarchical model for which the simplicial complex is defined by three sets of random vectors $\{A_t, C_t\}, \{A_t, S'_{t+1}\}$ and $\{C_t, S'_{t+1}\}$.

In order to calculate the value of this measure we apply the iterative scaling algorithm, as discussed in more detail in the context of the synergistic information in Section 4.5.3.

Furthermore, we calculate six of the measures discussed in the previous chapter. Each of these measures examines exactly one of the connections among the sensor, actuator and controller nodes. Since the measures are defined using the same concepts as before, they are named exactly like their corresponding counterparts in Section 4.5.

However, the agents have a slightly different architecture compared to the ones in the previous chapter, therefore we list the formulas for the measures below. In Figure 5.10, we see the graphical representation of the split models corresponding to these six measures summarized in one depiction. The name of a measure is printed next to the dashed connection that is removed in the split system of this measure.

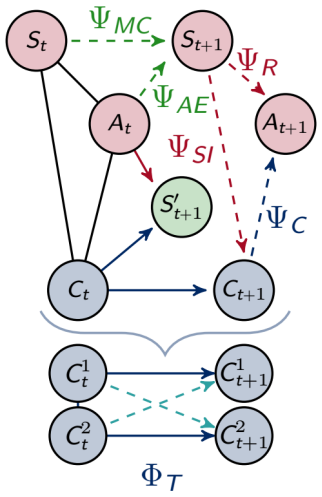


Figure 5.10: Graph depicting the split systems in case of Φ_T, Ψ_C, Ψ_{SI}, Ψ_R, Ψ_{AE} and Ψ_{MC}.

The first additional measure was initially introduced in Section 3.3 and can be seen in the context of Integrated Information Theory. It is called ground truth Integrated Information and results in

$$\Phi_T = \sum_j I(C^j_{t+1}; C^{J\setminus\{j\}}_t | C^j_t, S_{t+1}).$$

Here we only have two binary controller nodes, hence here $J = \{1, 2\}$.

In the context of our previous experiments we noted that the importance of the Integrated Information depends additionally on the information flow from and to the controller. Therefore, we additionally calculate the measures "control", Ψ_C, and "sensory information", Ψ_{SI}, which lead to the following conditional mutual information terms

$$\Psi_C = \sum_i I(A^i_{t+1}; C_{t+1} | S_{t+1}) \text{ and } \Psi_{SI} = \sum_j I(C^j_{t+1}; S_{t+1} | C_t).$$

In order to emphasize the importance of considering Φ_C and Φ_{SI} for the impact of the Integrated Information we additionally define in Section 4.6 the effective information integration:

$$\Phi_{EII} = \Phi_T \cdot \Phi_C \cdot \Phi_{SI}.$$

We calculate two measures regarding the empirical world model, $P(S_{t+1}|S_t, A_t)$. The first one quantifies the information flowing through the world and is used to asses Morphological Computation, as introduced in Section 4.5.3:

$$\Psi_{MC} = I(S_{t+1}; S_t | A_t).$$

In addition, we also consider the connection from the actuators to the next sensory state. The associated measure is called action effect:

$$\Psi_{AE} = I(S_{t+1}; A_t | S_t).$$

This quantifies the impact the actuators have on the next sensory state. Hence, it indicates whether the agents have an influence on their situation in the environment or not.

Lastly, we also discuss the reactive control Ψ_R, introduced in Section 4.5.4. This measure assesses the importance of the direct connection from the sensory to the actuator nodes:

$$\Psi_R = \sum_i I(A^i_{t+1}; S_{t+1} | C_{t+1}).$$

5.6 Results

In this section we discuss the results of our simulations. We use 1000 random input distributions for each sensor length of every type of agent, namely for the ideal, fully connected and split agents. All agents train for 20 000 steps and the measures are calculated for 90 different time points. More precisely, we apply the measures for the 9 time points listed below

$$50 \quad 100 \quad 200 \quad 500 \quad 1\,000 \quad 2\,000 \quad 5\,000 \quad 10\,000 \quad 20\,000$$

and 9 equidistant time points between each of them, as well as 9 equidistant time points between 0 and 50. Hence, the distance between the calculated measures is finer grained for smaller values and then becomes larger. This captures the more significant changes in the initial steps of the algorithm better compared to a completely equidistant approach.

Since the agents learn in real-time inside the racetrack and therefore experience very different sensory inputs depending on their individual behavior, there exist large variations in the results. Hence, in the following figures we only depict the arithmetic mean of a measure, depending on the sensor length and the time point.

Furthermore, we calculate the success rate of each agent by sampling how many time points the agent is stuck at a wall during the 20 000 training steps. Hence a success rate of 0.1 indicates that the agent was stuck 90% of the 20 000 training steps. Afterwards, we classify the third of the agents with the highest success rate as successful. In the case of the fully connected agents this leads to agents performing above 16.8% to be called successful while we refer to agents below 16.8% as unsuccessful. Dividing the agents there allows us to call only the agents successful for which the success rate increased significantly during learning.

We first consider a situation similar to the approach discussed in the previous chapter. Here we do not examine the learning process, but instead solely analyze the results for the controller complexity and Morphological Computation in the case of the successful, fully connected agents with an internal world model after 20 000 steps.

Note that in this case the controller complexity consists of the Integrated Information as well as the synergistic prediction. This is depicted in Figure 5.11 with the Morphological Computation measure Ψ_{MC} on the top, the Integrated Information measures Φ_T in the middle and the synergistic prediction Ψ_{SynP} on the bottom. There we observe that the controller complexity and the Morphological Computation behave antagonistically. Hence the results confirm the conclusion from the last chapter discussed in the context of Figure 4.27.

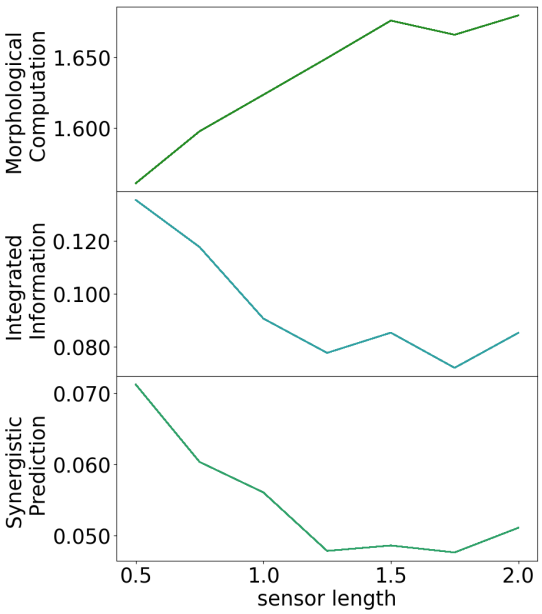


Figure 5.11: The results for Ψ_{MC}, Φ_T and Ψ_{SynP} after 20 000 steps of the fully connected agents with an internal world model.

When the sensors are too long, so that the agents almost always detect a wall, then this additional information is no longer beneficial for the agents and the Morphological Computation increases more moderately, while the Integrated Information and synergistic prediction values increase again.

As discussed in the introduction, Section 5.1, this leads to the question why agents with a well-adapted morphology would need a complex controller. There might be several reasons why a complex controller is necessary despite this relationship. Our experiments here might just be too simple so that more involved settings or tasks would lead to an agent requiring a complex controller. Here we argue that one reason for the necessity of a complex controller is that agents first have to learn how to interact with their environment, meaning they have to build their own world models.

5.6.1 The Ideal Agents

In order to examine this hypothesis further we first discuss the results for the ideal agents. These are the agents that do not have to learn an internal world model but have direct access to their empirical world model instead. Since the agents sample their empirical world models, while moving through their environment, the quality of this model is influenced by the behavior of the agents. We see the results for the different measures for the successful ideal agents in Figure 5.12. Note that these agents do not have an internal world model and therefore the controller complexity only consists of the Integrated Information.

We define successful agents as being the best thirty percent. For the ideal agents the best approx. 33% are the ones that have a success rate above $61,5\%$ and agents with a success rate below this are called unsuccessful. Hence, the ideal agents perform much better compared to the fully connected agents, which have to form an internal world model. The best third of the latter type of agents perform only better than 16.8%.

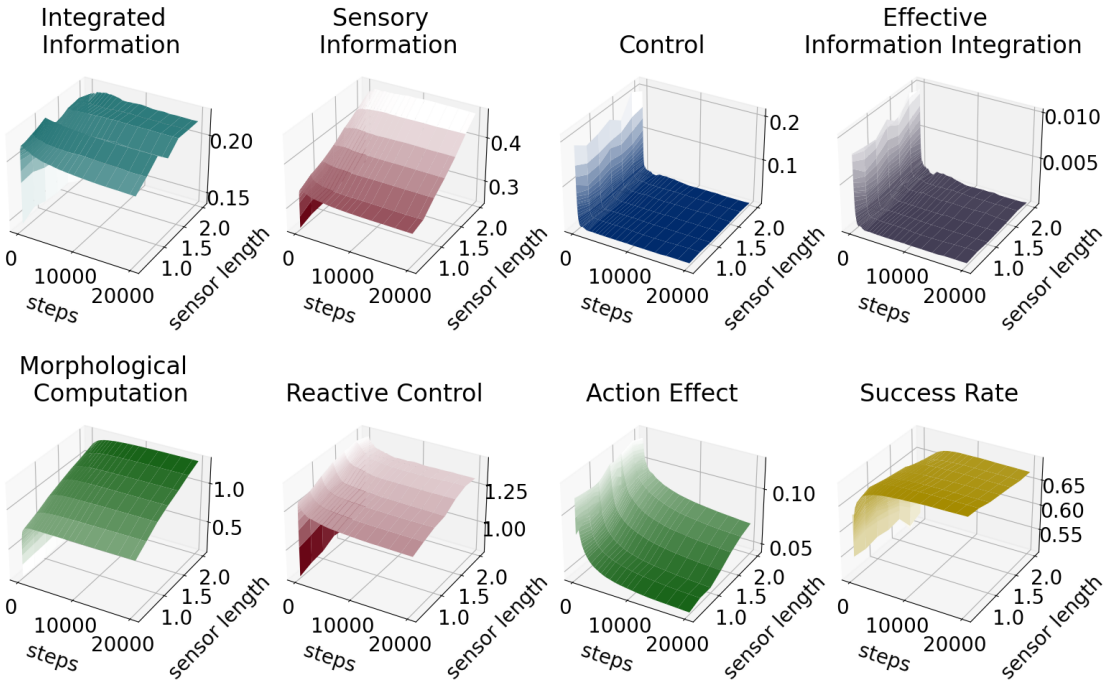


Figure 5.12: From the top left to the bottom right are the results for Φ_T, Ψ_{SI}, Ψ_C, Φ_{EII}, Ψ_{MC}, Ψ_R, Ψ_{AE} and the success rate in the case of the successful, ideal agents depicted.

On the top left in Figure 5.12, we see that the Integrated Information value seems to not change after the first few initial steps. In Section 4.6 we discuss that the importance of the Integrated Information additionally depends on the information flows to and from the controller, quantified by sensory information, Ψ_{SI}, and control, Ψ_C. While Ψ_{SI} increases with the sensor length, we can see the reason for the behavior of the Integrated Information in the results for control. After the first steps this measure is very close to 0 with an average value of 0.0021 at the 20 000th step. If $\Psi_C = 0$, then the Integrated Information value has no influence on the behavior of the agent at all. Since the controller has therefore also no impact on whether the agent reaches the goal or not, it is easy to check that then the information flow in the controller is not changed by the *em*-algorithm anymore. This leads to hardly any changes for the Integrated Information measure after the first steps.

That also explains the lack of a clear relationship between Morphological Computation and Integrated Information with respect to the sensor length here. The controller receives the sensory information from the environment and integrates it, but this integration process is not optimized with respect to the goal. Hence, the dynamics of the Integrated Information with respect to the sensor length does not reflect this antagonistic relationship. However, in the context of Figure 5.13, we are able to observe an asymmetric connection between Morphological Computation and Integrated Information.

Note that the modified *em*-algorithm also changes the information flow in the controller, even if the control value is zero. Therefore the above observations only hold for the ideal agents, because we apply the original *em*-algorithm here, not the modified one.

The effective information integration, Φ_{EII}, on the top right summarizes the behavior of Φ_T, Ψ_C and Ψ_{SI}. This is close to zero, which shows that there is next to no Integrated Information relevant for the behavior of the agent in this case.

The second row of Figure 5.12 depicts the Morphological Computation, reactive control, action effect and the success rate. Since these are the successful ideal agents, the success rate increases to around 0.7, meaning that these agents touch a wall only around 30% of the time. The Morphological Computation increases with the number of steps and the length of the sensors. Similarly, the reactive control value also increases with the sensor length, while the action effect decreases with the number of steps and the size of the sensors. Hence, the better the agents understand their environment, the smaller is the impact of their actions on their situation. We further examine the dynamics of the Morphological Computation, Integrated Information, reactive control and action effect in the context of Figure 5.13.

To summarize, in the case of the ideal agents the controller, Integrated Information included, is hardly needed in order to learn to perform a task. In fact, ideal agents without the ability to integrate information perform only slightly worse than fully connected ones. More precisely, the split ideal agents have an average success rate of 33.69% compared to the 33.83% in the case of the ideal agents that are able to integrate information. In this scenario, the success does not depend on the controller, but on the interaction of the agent with its environment and its direct reactions. We therefore now compare the Morphological Computation, action effect, reactive control and Integrated Information of successful and less successful ideal agents, depicted in Figure 5.13.

The first row of Figure 5.13 shows the results for the successful ideal agents. These agents have a much higher Morphological Computation over all compared to the unsuccessful agents in the second row. The Morphological Computation measures how much the next sensor state depends on the last sensor state given the action and it is calculated on the empirical world model. This means that the successful agents found strategies to move in their environment in a way such that the next point in time is more predictable, given the last sensor state, compared to the unsuccessful agents.

Looking at the results for reactive control we see that the successful ideal agents use

significantly less reactive control compared to the unsuccessful ones. Hence, if the next point in time is more predictable, then the need to immediately react to a sensory input is not as high. On the other hand, if the next sensory state is uncertain, then the agent has to rely more on its direct reactions to sensory stimuli. This is supported by the results of the action effect. The next sensory state of the successful agents does not rely as much on its action compared to the unsuccessful ones. So these reactive actions, that the agents have to perform in a more uncertain environment, have a higher impact on the next sensory state.

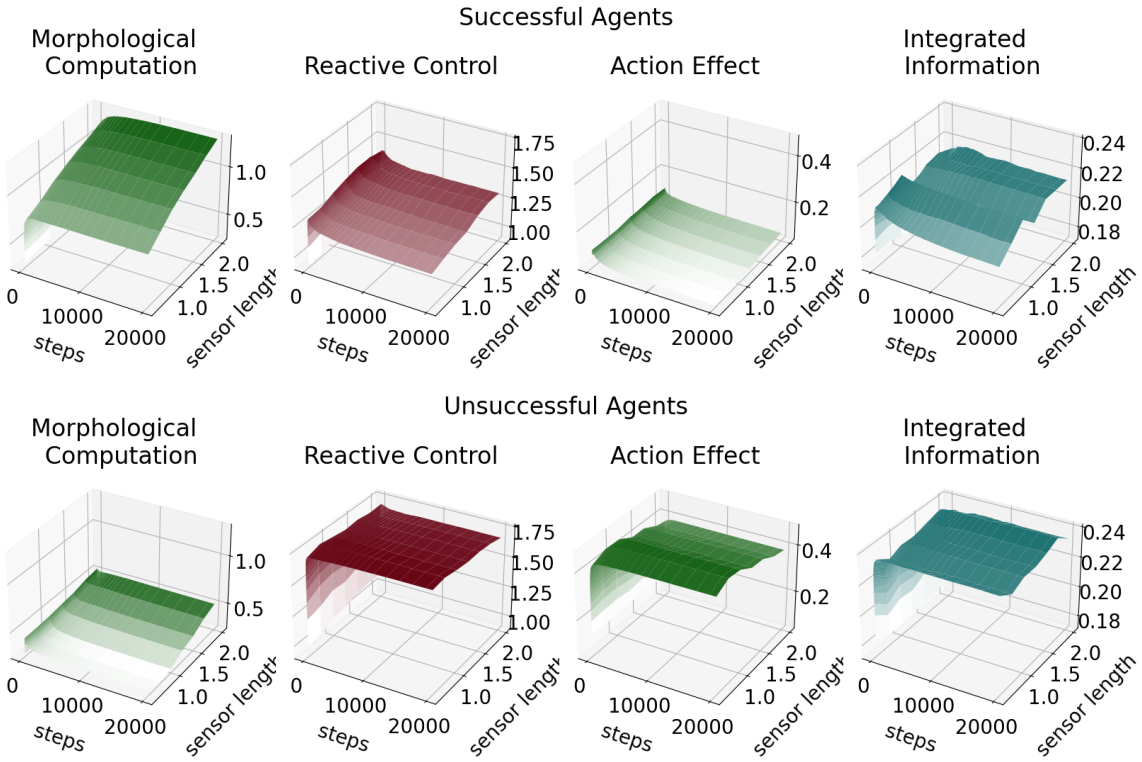


Figure 5.13: The results for Ψ_{MC}, Ψ_R, Ψ_{AE} and Φ_T for the successful ideal agents on the top and the unsuccessful ideal agents in the bottom row.

Similarly, the Integrated Information is overall higher in the case of the unsuccessful agents. There the agents have a lower Morphological Computation and we again observe an antagonistic relationship between these two concepts. Previously we noted that the Integrated Information is not influenced by the *em*-algorithm after the first few steps, however, the observation made here refers to the value that the algorithm reaches exactly during these first steps.

Note that this relationship is not reflected in the effective information integration since this value is the product of the Integrated Information, sensory information and control, and the controller has in both cases next to no influence on the behavior of the agents. The average value for the effective Integrated Information after 20 000 steps is 0.00017 for the successful and 0.00011 in case of the unsuccessful agents.

To conclude, ideal agents that have access to their correct world model have no need for a complex control architecture, a brain. Their success depends mostly on the interaction of their bodies with the environment, which is influenced by the sampled, empirical world model.

5.6.2 Varying the Accuracy of the World Model

In order to further examine this connection between the quality of the world model and the need for effective information integration, we additionally analyze agents that are only able to sample their empirical world model for a part of the total 20 000 steps. These agents move in their environment, sample the empirical world model and learn their behavior, exactly like the ideal agents up to a certain point. After that point the world model stays fixed and the agents then have to use this inaccurate world model to find the best behavior for the remainder of the 20 000 steps.

We distinguish between 9 different cases, namely agents that sample the world model for 50, 100, 200, 500, 1 000, 2 000, 5 000, 10 000 or the full 20 000 steps. The last type of agents, which sample the empirical world model for 20 000 steps, are exactly the ideal agents discussed in the previous section. Here we use the same threshold of $61,5\%$ as for ideal agents for all agents to classify them as successful or unsuccessful. This allows us to compare the percentages of successful agents across the different types of world models. In Table 5.1 we list the average success rate and the percentage of successful agents for the different world models after 20 000 steps.

world model sampling steps	average success rate	percentage of successful agents
50	27.7%	18.0%
100	30.1%	26.7%
200	30.9%	31.2%
500	31.6%	34.6%
1 000	31.9%	34.4%
2 000	31.9%	34.7%
5 000	31.2%	33.2%
10 000	31.6%	33.9%
20 000	31.7%	33.8%

Table 5.1: The average success rate and the percentage of successful agents after 20 000 steps for the different world models, rounded to the first decimal place.

There we observe that both values increase with the number of steps the world model is trained for the first four world models. This behavior is expected, because it shows that the success of the agents increases with the accuracy of their world model. The agents with world models trained for 1 000 steps or more are very similar in their average success rate and the percentage of successful agents varies only between 33.2 and 34.7. This leads to the conclusion that the agents need around 1 000 steps to learn an accurate world model for this setting. Hence, the different Integrated Information measures have very similar results for the agents that learn their world model for 1 000 steps or longer. Therefore, we only plot the results of the different measures up to the models learning for 2 000 steps, which increases the visibility of the differences between models with less accurate world models.

Figure 5.14 depicts these results for the different measures with respect to the number of steps and the accuracy of the world model. In this case we display the arithmetic mean over the different sensor lengths. On the bottom right we see that these agents have an average success rate of around 67.5 % after the 20 000 training steps.

The Integrated Information and sensory information increase with the number of steps and vary only very slightly with respect to the accuracy of the world model. On the other hand, the value for control decreases with the number of steps and results in a higher value in case of a world model trained for only a few steps. As discussed in the context of Figure 5.12, a small control value indicates that the Integrated Information is hardly affected by the *em*-algorithm and therefore does not change with the number of steps or the accuracy of the world model. The effective information integration exhibits a dynamic similar to control.

The value for Morphological Computation, depicted in the bottom row on the left, displays the opposite dynamics. This value increases with the number of steps and the accuracy of the word model. Hence, the better the agent can model the dynamics of the world, the better its body can interact with the environment.

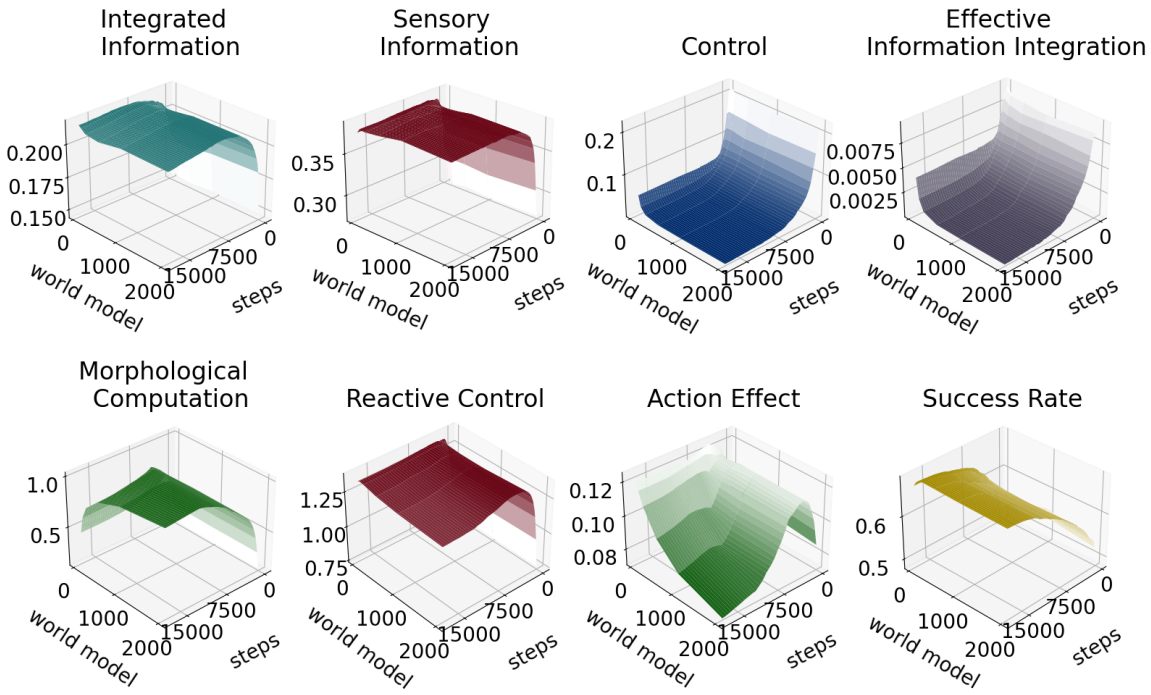


Figure 5.14: From the top left to the bottom right are the results for the successful agents with the different types of world models for Φ_T, Ψ_{SI}, Ψ_C, Φ_{EII}, Ψ_{MC}, Ψ_R, Ψ_{AE} and the success rate.

The value measuring reactive control is almost constant after the first 5 000 steps during which the action effect exhibits an increase and then decrease regardless of the world model. Interestingly, the action effect then declines asymmetrically with respect to the accuracy of the world model. So, the more accurate the world model is, the more the agent can interact with its environment using Morphological Computation, which reduces the influence of the agent's actions on its situation.

In Figure 5.15 we highlight the relationship between Morphological Computation, on the top, and effective information integration, displayed on the bottom, with respect to the accuracy of the world model. There we display only the arithmetic mean over the different sensor length after 20 000 steps.

While the Morphological Computation increases with the accuracy of the world model, the effective information integration decreases. In the introduction, Section 5.1, we motivate

the intuition behind these concepts using the example of a child trying to ride a bike. The better the child understands the dynamics of its environment, the more it can make use of them and drive faster to stabilize the bike.

Hence, a more accurate world model leads to a higher Morphological Computation value, which in turn reduces the necessity for the controller to integrate information.

This concludes the analysis of the accuracy of the empirical world model in relationship to the information flow inside the ideal agents. In the next section we discuss the fully connected agents that have an internal world model.

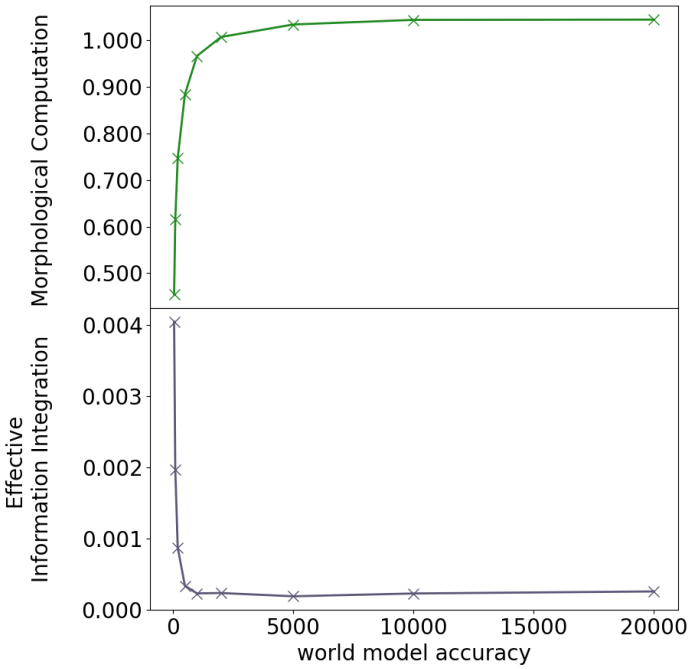


Figure 5.15: The results for Φ_{EII} and Ψ_{MC} for the different world models after 20 000 steps

5.6.3 The fully connected Agents with an internal World Model

Agents are rarely able to fully grasps all the relevant aspects of their environment and instead have to build a more limited, internal world model. Therefore, we discuss in this section the results for the fully connected agents that optimize their internal world model via the modified *em*-algorithm defined in Section 5.4.1.

First, we consider the results for the controller complexity in case of the successful and unsuccessful agents. The complexity of the controller consists of the Integrated Information and synergistic prediction. Here, the threshold for success lies at 16.8%. The two top rows of Figure 5.16 depict on the left the success rate of the fully connected agents. In the middle there is the Integrated Information for these agents and the graphs on the right show the results for the synergistic prediction. Here the first row consists of the 3-dimensional results w.r.t. to the time steps and the sensor length and in the second row we have a 2-dimensional depiction where the x-axis displays the time steps.

The unsuccessful agents have an average success rate of approximately 10% after 20 000 steps. Their Integrated Information value lies between 0.3 and 0.4 and their synergistic prediction value is between 0.1 and 0.15 after the first 2 000 steps. There is no significant decrease or increase in the Integrated Information or synergistic prediction with respect to the number of steps and the dynamics in relation to the sensor length has an overall downward direction.

Now we compare these results to the controller complexity of the successful agents. Similarly to the top rows, the two rows on the bottom of Figure 5.16 show the success rate of the fully connected, successful agents on the left, the Integrated Information in the middle and the synergistic prediction on the right. The average success rate of these agents after 20 000 steps is approx. 25%. Since we sample the success rate, it can be very high after the first few steps, which results in the first steep decline around 500 steps.

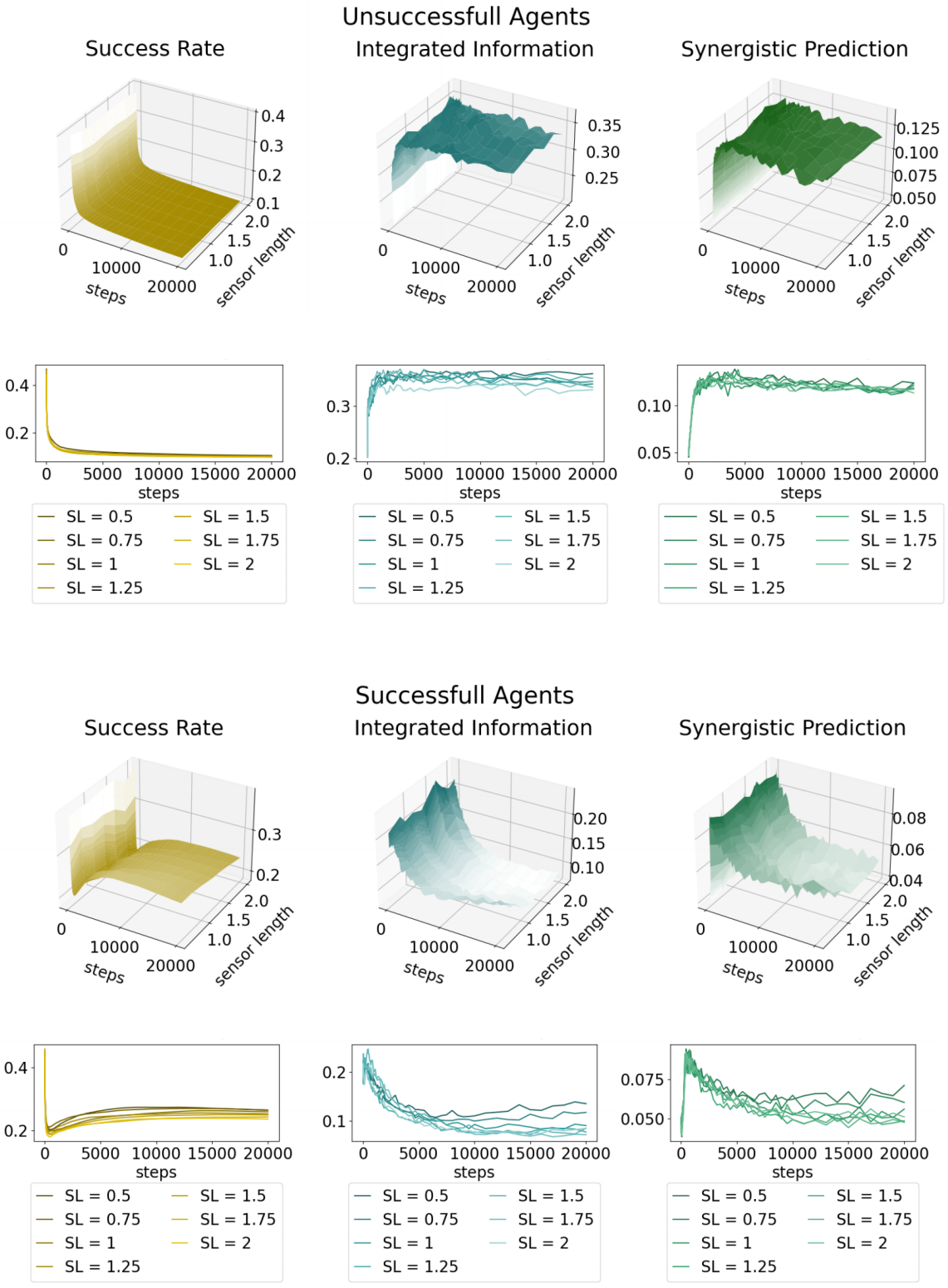


Figure 5.16: The two top rows consist of the success rate, Integrated Information and synergistic prediction results for the unsuccessful, fully connected agents and the two bottom rows depict the same results for the successful agents. Here SL stands for sensor length.

Next to the success rate we observe that the Integrated Information value decreases with the number of steps. It reaches a value roughly between 0.05 and 0.15 after 20 000 steps. The synergistic prediction exhibits a similar dynamic leading to a value between 0.04 and 0.07. Hence in the case of the successful agents the Integrated Information and synergistic prediction values go down to a significantly lower level compared to the unsuccessful agents. Additionally, in the last row we observe an increase of the Integrated Information and synergistic prediction values in the first approx. 500 steps.

Before we interpret these results in connection to the learning behavior of the agents, we first discuss the values for the measures Ψ_C, Ψ_{SI}. These measures give insights to the effect the information integration has on the action of the agent and combined lead to the definition of Φ_{EII}, described in Definition 20. In Figure 5.17 we depict these three measures, namely control, sensory information and effective information integration, in the case of the successful and unsuccessful agents.

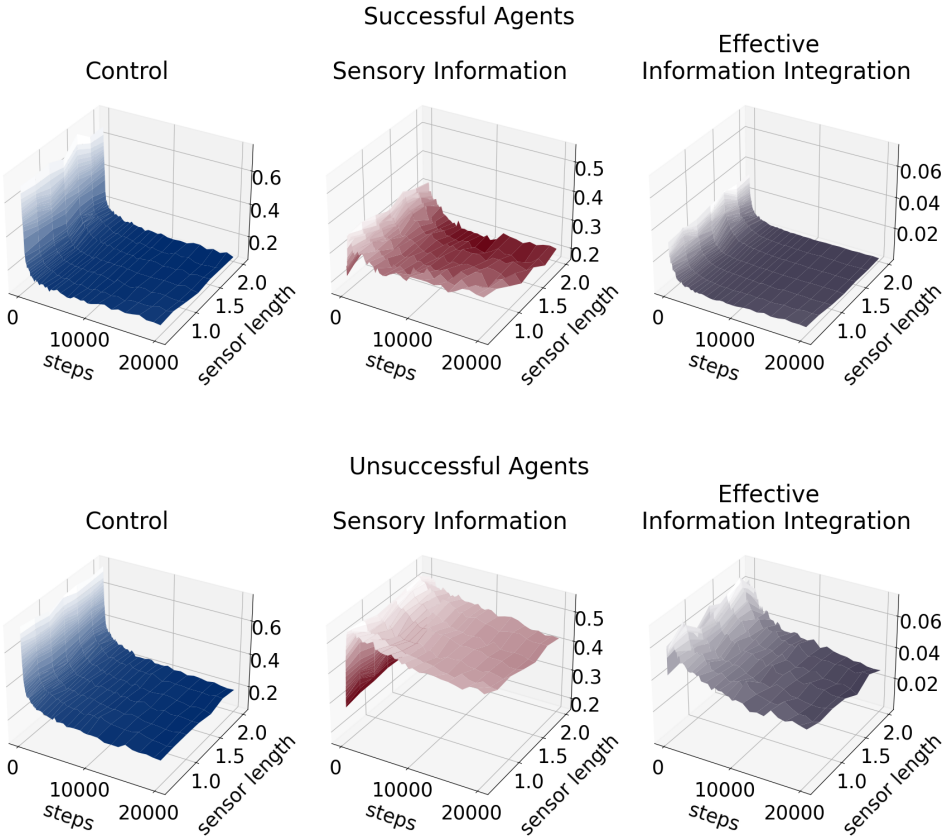


Figure 5.17: The measures Ψ_C, Ψ_{SI} and Φ_{EII} for the successful, fully connected agents in the top row and for the unsuccessful agents in the bottom row.

In the first row of Figure 5.17 we see the results for the successful agents. The measures Ψ_C and Ψ_{SI} behave similarly compared to the Integrated Information and synergistic prediction values of the successful agents. They decrease with the number of steps and with the length of the sensors, although this is not visible in the figure for Ψ_C. Therefore, the control values, rounded to the third decimal place, for the different sensor length after 20 000 steps are listed in the Table 5.2.

It is important to notice that Ψ_C is not zero here but has in average a value of approx. 0.1. So the controller has a small, but existing, influence on the behavior of the agent. The ideal

agents, discussed in the previous section, have an average control value after 20 000 steps of 0.0021. Comparing the ideal agents with the results discussed in this section indicates that the controller has a larger impact on the action of the agents when they have to form their own internal world models.

sensor length	0.5	0.75	1	1.25	1.5	1.75	2
Ψ_C	0.118	0.124	0.101	0.082	0.094	0.089	0.085

Table 5.2: The results for Ψ_C in case of the successful fully connected agents after 20 000 steps.

The influence of the controller is even larger in the case of the unsuccessful agents, depicted in the bottom row of Figure 5.17. There all of the three measures, namely Ψ_C, Ψ_{SI} and Φ_{EII}, display a similar dynamic compared to the successful agents, but with an overall higher value. The decrease of the effective information integration is flatter for the unsuccessful agents.

To summarize, the Integrated Information value is higher in the case of unsuccessful agents, compared to the successful ones. This holds also true for the measures related to the importance of the Integrated Information for the behavior of the agent, namely control, sensory information and effective information integration.

Now we combine these results with the insights of the previous section. There we vary the accuracy of the world model of an agent and observe that the effective information integration decreases when the accuracy of the world model increases. Therefore, a high controller complexity could be important as long as the agents have not been able to learn the correct world model. Without a correct world model the agents are not able to find actions that would allow them to optimally use their interaction with the environment. Hence, the agents first have to learn an internal world model and are then able to find a strategy that leads to a high Morphological Computation value.

This explanation would fit to the first increase in the controller complexity, given by Integrated Information and synergistic prediction, with respect to the step size for the successful agents. During the first steps the agents still have to learn the correct world model. If they are successful in learning this model, then the Morphological Computation should be high leading to a decrease in the Integrated Information and synergistic prediction.

In order to further analyze this mechanism we depict in Figure 5.18 the Morphological Computation, reactive control and action effect for the successful agents in the first and unsuccessful agents in the second row.

There we can confirm that the successful agents have a higher Morphological Computation value with an average of approximately 1.63 compared to the unsuccessful ones that have an average of roughly 1.5 after the 20 000 steps. Additionally, the action effect in the last column of Figure 5.18 displays the opposite dynamics. The higher the Morphological Computation value, the less influence the agents have on their next sensory state. This is in line with the results of these measures for the ideal agents, discussed in connection with Figure 5.13.

Interestingly, the reactive control displays a different dynamic compared to the reactive control of the ideal agents. The ideal agents have access to their empirical world model whereas the fully connected agents, discussed here, form an internal world model. This internal world model seems to limit the agents to strategies that use less reactive control, rather than more.

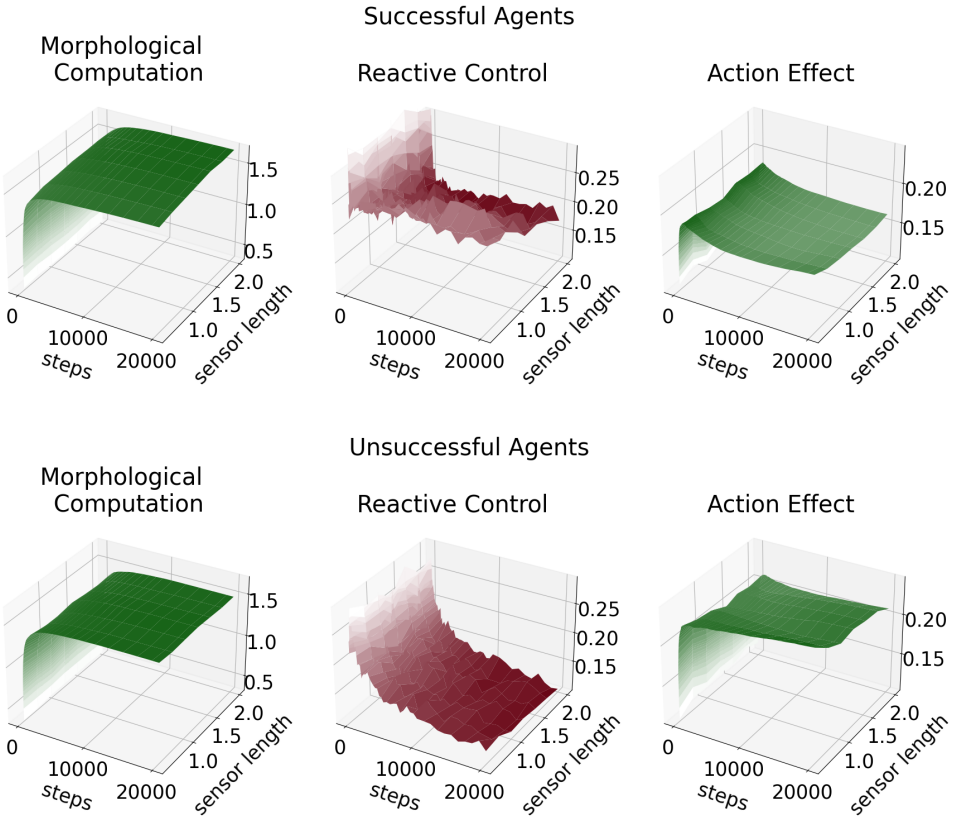


Figure 5.18: The results for Morphological Computation, reactive control and action effect for the successful and unsuccessful fully connected agents.

In order to understand this internal world model better we now discuss four measures of the information flow inside it, namely the full prediction Ψ_{FP}, actuator prediction Ψ_{AP}, controller prediction Ψ_{CP} and synergistic prediction Ψ_{SynP}. These are defined in Section 5.5. Since the ideal agents do not have an internal world model, we cannot compare the results of these measures to the previous section. Figure 5.19 depicts them for the successful agents in the first and for the unsuccessful agents in the second row.

The full prediction stays fixed after the first few steps of the agent and increases with the length of the sensors. Hence, the additional information from the longer sensors lead to a better predictability of the next sensory state. This is not the case for the unsuccessful agents. There the value for the full prediction is exactly the opposite, it decreases when the sensor length increases. So, for the unsuccessful agents there is a stronger dependency of the next sensory state on the last sensory state and action for shorter sensors, even though they do not provide as much information about the environment as the longer ones.

Furthermore, the actuator and synergistic prediction decrease with the number of steps and the increasing sensor length for the successful agents. For longer sensors the prediction of the next sensory state does not depend as much on the chosen action or interaction between actuator and controller as for shorter sensors. Instead, there the agent relies more on the controller states alone, which we can see in the overall high value for the controller prediction and the slight increase in this value with the number of steps.

The unsuccessful agents, on the other hand, have much higher values for the actuator and synergistic prediction, while the controller prediction decreases with the number of steps taken. Hence, the world models of the unsuccessful agents depend on the actuator and the controller states as well as their interaction.

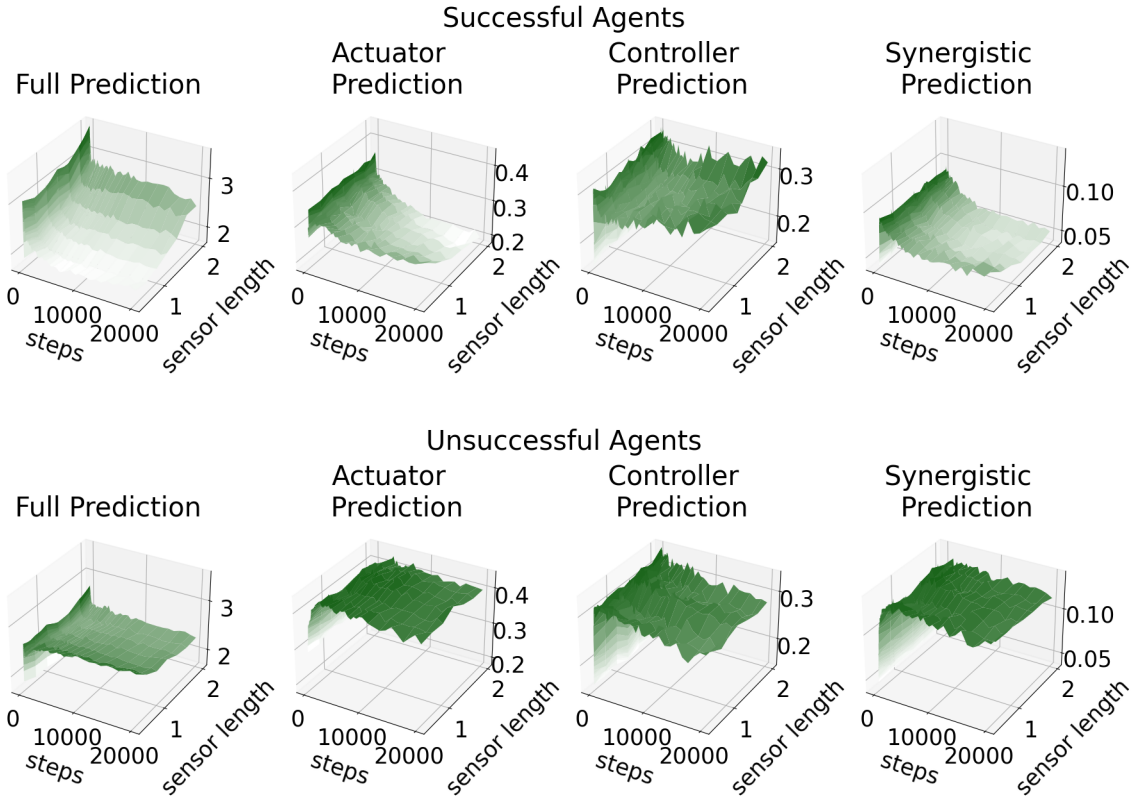


Figure 5.19: Results for the measures $\Psi_{FP}, \Psi_{AP}, \Psi_{CP}$ and Ψ_{SynP} in case of the successful agents in the top row and for the unsuccessful in the bottom row.

All in all, the successful agents converge to a strategy for which the prediction of the next sensory state depends not as much on the selected action and the interaction between controller and actuator, but more on the controller values alone. Hence, the mechanisms inside the controller are vital for the prediction process in the case of the successful agents. This indicates that information integrated in the controller might be crucial for a good world model. Therefore we hypothesize that a high Integrated Information facilitates the learning of an accurate world model.

However, is the high Integrated Information that we observe in Figure 5.16 even necessary for learning or is it just a byproduct of the learning algorithm? This value might simply be a result of the dynamics of the Morphological Computation and their relationship without having an impact on the leaning on its own.

To clarify the importance of the role of the Integrated Information for learning a world model and a strategy we next discuss the results of the split agents. These agents are not able to integrate information in their controllers.

5.6.4 The split Agents

The split agents, discussed in this section, have to learn an internal world model but are not able to send information from one controller node to the other and therefore they have a fixed Integrated Information value of zero. Consequently, also the effective information integration is zero for these agents. Hence, the controller complexity solely depends on the synergistic prediction, calculated on the internal world model, in this case.

In this section we divide the split agents in successful and unsuccessful ones by applying the same criterion as in the case of the fully connected agents, which is 16.8%. Although the split agents perform significantly worse compared to the fully connected ones, so that

this does not lead to a classification of the best third, it nonetheless allows us to directly compare fully connected and split agents with a similar success rate.

Before we analyze the results of the different information theoretic measures we first discuss the overall performance of the split agents. Table 5.3 lists the average success rates after 20 000 steps of fully connected and split agents. In addition we compare these rates with the average success rate of 1 000 agents that each perform 20 000 random movements inside the racetrack.

	random movement	fully connected agents	split agents
average success rate	$\approx 7.95\%$	$\approx 15.21\%$	$\approx 8.01\%$

Table 5.3: Arithmetic mean of the success rates after 20 000 steps of the agents with random movement and the fully connected and split ones.

There we observe that the split agents perform in average barely better than the agents that move randomly. The fully connected agents, on the other hand, are almost twice as successful as the split ones. Furthermore, there is a considerable difference between the number of successful split agents compared to fully connected ones. Only approx. 2.1% of split agents are successful, opposed to roughly 33.3% of fully connected ones.

In summary, the split agents perform only marginally better than agents that move purely at random and only very few split agents are successful. This strongly supports the hypothesis that the ability to integrate information in the controller is necessary for learning a correct world model.

Now we discuss the results of the measures in case of the successful split agents. They are depicted in Figure 5.20. The sensory information increases with the number of steps and decreases for longer sensors. It is overall higher compared to the sensory information of the successful fully coupled agents. The Φ_C behaves similarly in both cases, but for the successful split agents the control value after 20 000 steps is lower, in average around 0.08. So, without information integration in the controller the information flow coming in is more important while the commands sent from the controller has less impact.

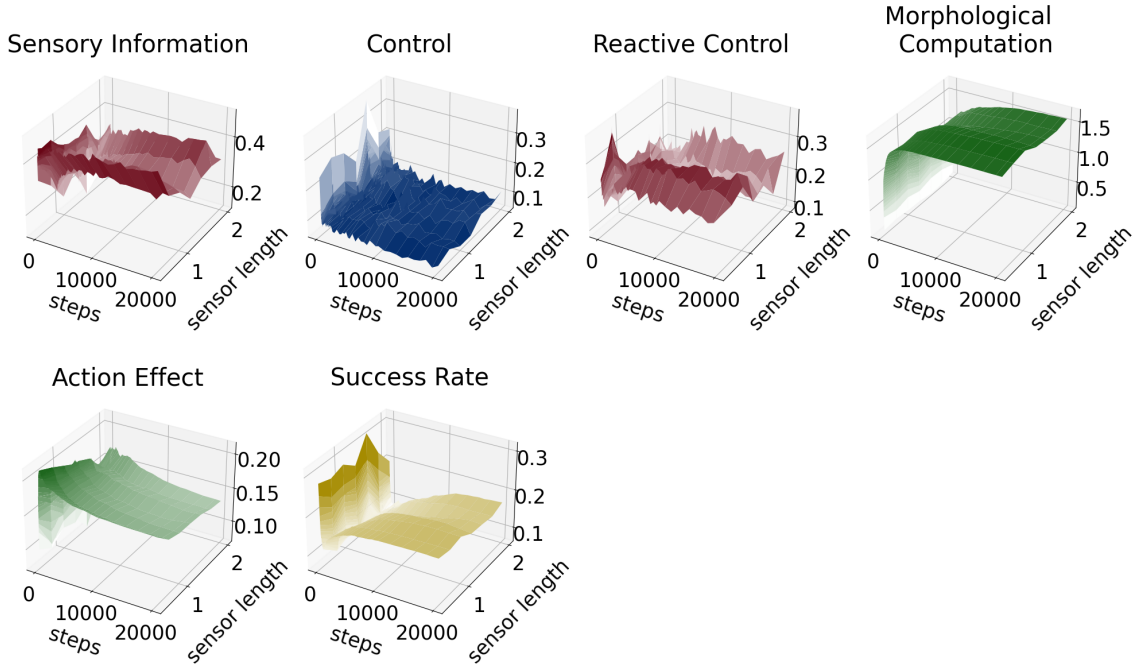


Figure 5.20: Results of the measures Ψ_{SI}, Ψ_C, Ψ_R, Ψ_{MC}, Ψ_{AE} and the success rate for the successful split agents.

The reactive control values are slightly higher, but exhibit the same decreasing dynamic with increasing sensor length as the reactive control values of the fully connected agents. On the other hand, the Morphological Computation and action effect values are lower for the split agents. Hence the successful split agents react more directly to their environment with a a slightly lower Morphological Computation and action effect.

Note that we only display the results of a small number of agents here since only around 2.1% of the total 1 000 split agents are successful. Therefore, we do not have enough data to draw any further conclusions from the small differences between the split and the fully connected agents with an internal world model, which we discuss here in the context of Figure 5.20.

However, we observe a significant difference between the split and fully connected agents by considering the structure of their internal world models next. There we calculate the measures for the full prediction Ψ_{FP}, actuator prediction Ψ_{AP}, controller prediction Ψ_{CP} and synergistic prediction Ψ_{SynP}. The results in case of the successful split agents are depicted in Figure 5.21 in the top row and for the unsuccessful agents in the bottom row.

The results for the unsuccessful split agents are very similar to the results of the unsuccessful fully connected agents. These are depicted in the bottom row of Figure 5.19. This might be the case, because the unsuccessful fully connected and split agents have an equally inaccurate world model with similar internal information flows.

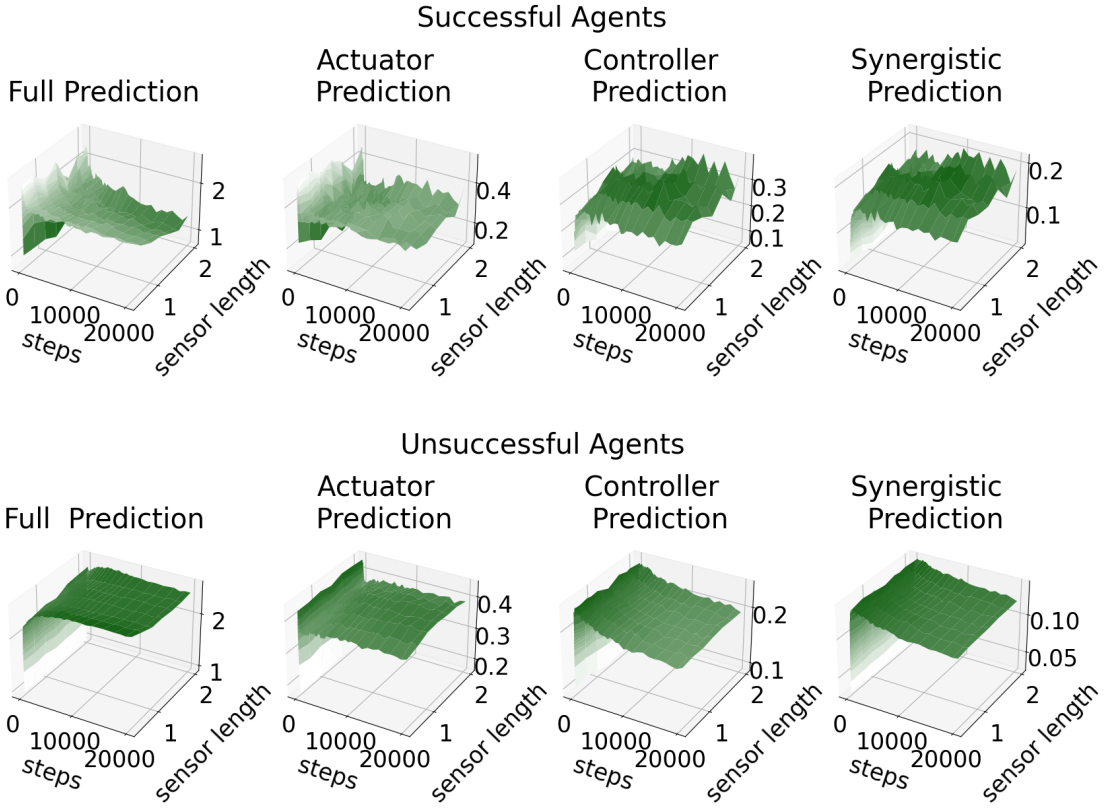


Figure 5.21: Results of the measures Ψ_{FP}, Ψ_{AP}, Ψ_{CP} and Ψ_{SynP} for the successful split agents in the top and for the unsuccessful split agents in the bottom.

For the remainder of this section we now focus on the results of the successful agents. The full prediction is overall lower for the split agents compared to the fully connected ones and it decreases with increasing sensor length. So, the next sensory state is in general not as predictable compared to the fully connected agents. The actuator and controller predictions exhibit similar dynamics to the successful fully coupled agents.

The main difference between the world models of the successful fully connected and successful split agents lies in the synergistic prediction. It is much higher for the successful agents that are not able to integrate information. This leads to the conclusion that for these split agents the internal world model, and therefore the prediction process, has to combine information and become much more complex. The fully connected agents are able to integrate the information in their controllers and do not need such a complicated world model.

This concludes the discussion of the experiments with learning agents. In the next section we summarize all the insights about the relationship between Morphological Computation and Integrated Information that we gained by analyzing the acting agents in Chapter 4 in addition to the ideal, fully connected and split agents in this chapter.

5.7 Summary and Discussion of Chapter 5

In this chapter we discuss the dynamics of the information flows in learning, embodied artificial agents. Here we differentiate between agents that have an internal world model and ones that have a direct access to their empirical world model. The latter agents are trained using the *em*-algorithm, as in Chapter 4, whereas we use an adapted *em*-algorithm for the agents with an internal world model. This adapted algorithm alternates between optimizing the behavior in order to reach a goal and updating the internal world model.

The agents move inside a racetrack and learn to not touch the walls, similarly to the agents in the previous chapter. However, here we adapted the movements and the goal slightly to increase the influence the controller has on the actions of the agents. Additionally, the agents discussed in this chapter are actually trained after each step they take inside their environments. Hence, the empirical world model of each agent depends on the actions they perform.

The outcomes of our experiments regarding Integrated Information and Morphological Computation support the results discussed in Chapter 4. In particular, we again observe the antagonistic relationship between these two measures. The previous results led to the insinuation that agents with a highly adapted morphology could have no use for a complex control architecture with a high Integrated Information value at all. There are many possible ways to address this issue. One possibility is that our tasks are simply too easy, so that an agent truly only needs Morphological Computation in order to be successful.

Despite the simplicity of our example, we are able to offer an additional solution to the posed problem in this chapter. We theorize that learning to predict the environment, meaning forming a correct internal world model, results in a necessity for a complex controller. Ideal agents, which do not have to learn to predict their environment, do not require a complex controller at all in order to learn to be successful at the task. The agents with an internal world model, on the other hand, do. Comparing these fully connected agents with the split ones, which are not able to integrate information in their controllers, leads to the observation that the split ones are not able to predict the next sensory state well. The split agents perform in average only marginally better than completely random moving agents and there is only a very small percentage of split agents that are successful.

The controller complexity of these fully connected agents consists of the Integrated Information and the complexity of the internal world model, measured by the synergistic

prediction. For the successful agents the controller complexity is first high, while they adapt their world model to their environment, and then it decreases. We argue that this decrease could result from a rise in Morphological Computation that is facilitated by the correct world model. This is supported by the results of the Morphological Computation measure, which is higher in the case of the successful agents.

Additionally we observe that for successful split agents with an internal world model the process of predicting becomes more complex. The complexity of the internal world model is much higher for the successful split agents compared to the successful fully connected agents. Hence, without an information integration among the controller nodes the internal world model itself needs to combine the information from the controller and the actuator in order to form an accurate prediction. This again supports the claim that an agent needs to integrate its available information in order to build an accurate internal world model.

One caveat of this analysis is that only a very limited number of split agents with an internal world model are successful. These agents tend to get stuck at a wall and are then not able to move away. The reason for this behavior might be that they simply learned to activate the "fast forward" action of a wheel when the sensor on the same side detects a wall. As long as the agent moves around and only one sensor detects a wall at a time this is a valid strategy. However, as soon as the agent is stuck and both sensors detect a wall the agent now activates both wheels with the fast movement and is therefore not able to turn. Hence, although this experiment is already very simplistic, it might be too complicated for small split agents with an internal world model.

A final discussion of the results and an outlook on potential future research is given in Chapter 7. In Section 7.2.1 we in particular summarize the results from Chapter 4 and Chapter 5 regarding the relationship between the Integrated Information and Morphological Computation.

6 Advanced Theoretical Results

In the following two sections we discuss further theoretical results regarding two information geometric algorithms that we apply in the previous chapters. First, we address the case in which the *em*-algorithm is used to find the MLE of a model with latent variables, as described in Section 2.5.1. This was discussed in Section 3.4.1 as a method to calculate the Integrated Information measures Φ_{CII}. Here we define three additional methods to increase the state space of the latent variable, also called hidden variable, that also take a previously found local minimum into account.

Secondly, in Section 6.2, we discuss the iterative scaling algorithm that we introduced in Section 2.5.2. Here we define a condition under which this algorithm converges in only one cycle. The exponential families defined in this section include hierarchical models and this condition is a generalization of the known "Running Intersection Property".

6.1 Gradually Increasing the Latent Space in the *em*-algorithm

The *em*-algorithm can be used to find the MLE of a model with only partially observed data, as described in Section 2.5.1. It iterates between an e-projection to a set $\mathcal{M}_1^\ell$ and an m-projection to a set $\mathcal{M}_2^\ell$. In the case of partially observed data the random vector can be divided into hidden and known variables with the state spaces $\mathcal{Y}_h$ and $\mathcal{Y}_k$, respectively. The set $\mathcal{M}_1^\ell$ then consists of the probability distributions P for which the marginals on the known variables coincide with the empirical distribution of the observed data, as defined in Equation (2.4). Applying the *em*-algorithm for $|\mathcal{Y}_h^\ell| = \ell$ leads to

$$D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_i^\ell \parallel Q_i^\ell) \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_i^\ell \parallel Q_{i+1}^\ell) \geq \cdots \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell),$$

with $P_i^\ell, P_\star^\ell \in \mathcal{M}_1^\ell$ and $Q_i^\ell, Q_{i+1}^\ell, Q_\star^\ell \in \mathcal{M}_2^\ell$. The question remains how to choose ℓ, the dimension of the latent space.

In Section 3.4.1 we use this form of the *em*-algorithm in order to calculate the Integrated Information measure Φ_{CII}. This measure accounts for an unknown exterior influence for which also the size of this exterior influence is undefined. One method to handle the missing size of the latent space is to calculate the measure for increasing latent spaces independent of each other, as depicted in Figure 6.1.

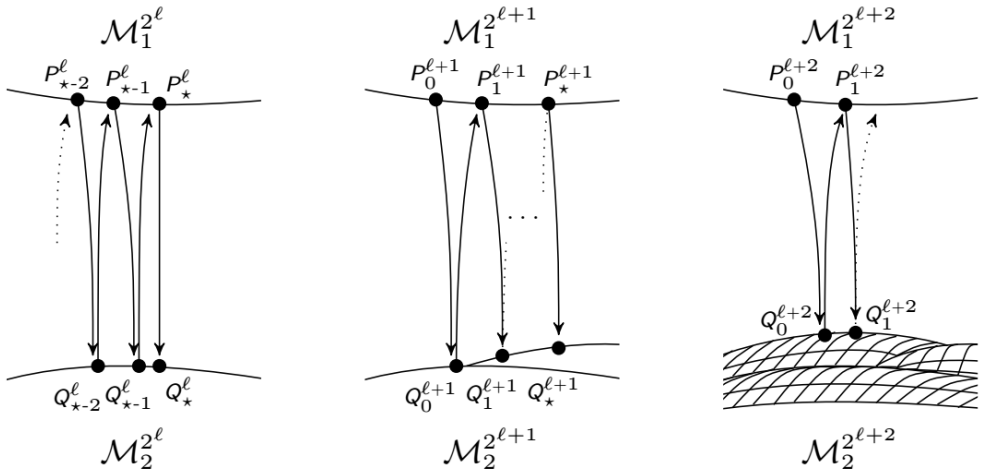


Figure 6.1: Sketch of the traditional method of applying the *em*-algorithm to different sized latent spaces.

Afterwards we can take the minimum of the resulting values.

We apply this approach in Example 3. There we use a system with three binary variables, for which the transition probability is defined by the connection matrix V with

$$V = \begin{pmatrix} -0.5 & 0.01 & -1 \\ 0.01 & 0 & 0.3333 \\ 0 & 0 & 0 \end{pmatrix}.$$

The system is inspired by the Boltzmann machine and an introduction to this setting is given in Section 2.1.

Two of the binary variables are considered to be known and the third one is the exterior influence. Since this influence should be unknown, we apply the *em*-algorithm to the marginalized stationary distribution.

In Figure 6.2 we depict all the local minima of 5 random input distributions for latent state spaces with the sizes 2, 4, 8 and 16. The inverse temperature β ranges from 0.01 to 10 with a step size of 0.01 and the algorithm performs 1000 steps for each β

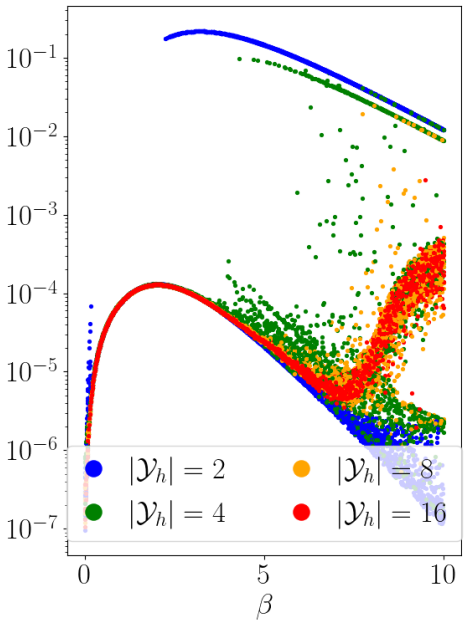


Figure 6.2: The results of the traditional method for five random initial distributions.

and each size of the state space. In this example we observe that for $\beta > 7.5$ the local minima found for the state spaces with size 2 and 4 are in most cases better compared to the ones in the larger dimensional spaces. On the other hand, for $\beta < 5$ the best results are reached in case of $\ell = 16$.

Hence, by using a larger number of random input distributions and different sized latent spaces this way we are able to gain increasingly better local minima. We refer to this approach as the "traditional" method and we applied this in Chapter 3.

However, this method does not use the information about the local minimum from the lower dimensional space in order to find a better solution in a higher dimensional latent space. Instead it starts with a different random distribution each time. In order to improve this approach we now describe three alternative methods, namely the "natural", "safe" and "experimental" method. These all make use of the local minima calculated in the smaller space in order to gain a new initial distribution for the larger space.

6.1.1 Natural Method

We name the first approach the "natural" method. Here we increase the state space of the hidden variable by an additional state $\bar{y}_h$. Note that we can also use this same method to add multiple additional states in each step but to increase readability we introduce only one additional hidden state in the definitions below.

Let $P_\star^\ell$ and $Q_\star^\ell$ be the distributions in $\mathcal{M}_1^\ell$ and $\mathcal{M}_2^\ell$, respectively, that the *em*-algorithm arrived at. These do not have to be fixed points of the algorithm, meaning that the algorithm can be stopped before it reaches convergence.

A natural way to project to the next larger latent space is to set the probability of the new state $\bar{y}_h$ to 0. Hence, we can define a distribution in the larger space in the following way

$$P_0^{\ell+1}(y_k, y_h) = (1 - \delta_{\bar{y}_h}(y_h)) P_\star^\ell(y_k, y_h), \quad \forall (y_k, y_h) \in \mathcal{Y}_k \times \mathcal{Y}_h,$$

with the indicator function

$$\delta_{\bar{y}_h}(y_h) = \begin{cases} 1 & \text{if } y_h = \bar{y}_h \\ 0 & \text{otherwise} \end{cases}.$$

In the case of multiple new hidden states the function above equals one if and only if y_h lies in the set $\bar{\mathcal{Y}}_h$ that consists of the additional hidden states. Therefor this can be written as

$$\delta_{\bar{\mathcal{Y}}_h}(y_h) = \begin{cases} 1 & \text{if } y_h \in \bar{\mathcal{Y}}_h \\ 0 & \text{otherwise} \end{cases}.$$

The new distribution $P_0^{\ell+1}(y_k, y_h)$ is an element of $\mathcal{M}_1^{\ell+1}$.

Unfortunately, using this newly defined $P_0^{\ell+1}(y_k, y_h)$ as initial distribution for the *em*-algorithm does not lead to an improved local minimum in the larger space in most cases. One key point in the proof of Theorem 2.5.1 is that the *e*-projection of an element Q to $\mathcal{M}_1^{\ell+1}$ does not alter the conditional distribution of the hidden variables given the known ones $Q(Y_h|Y_k)$. Therefore the probability of $\bar{y}_h$ stays 0 under this projection.

The *m*-projection varies with the structure of $\mathcal{M}_2^{\ell+1}$, but it is easy to see that the probability of the state $\bar{y}_h$ remains 0, if $\mathcal{M}_2^{\ell+1}$ is a graphical model corresponding to a DAG. This is also the case for our example, where we consider a chain graph associated with the measure Φ_{CII}.
There the *m*-projection is defined in Proposition 7 in the following way

$$Q(b, w) = P(c_t) \prod_{j=1}^{n} P(c_{t+1}^j | c_t^j, w) P(w)$$

for all $(b, w) \in \mathcal{B} \times \mathcal{W}_C^m$ where $b = (c_t, c_{t+1})$ are known states and the w's are hidden. If for one $\bar{w}$ the probability $P(\bar{w})$ equals zero, then $Q(b, \bar{w}) = 0$. Therefore this approach does not lead to an improved minimum for the Integrated Information measure Φ_{CII}.

Instead, we define an intermediate step between $\mathcal{M}_2^{\ell}$ and $\mathcal{M}_1^{\ell+1}$ in order to perturb the system slightly by a small constant c

$$\hat{Q}^{\ell+1}(y_k, y_h) = (1 - \delta_{\bar{y}_h}(y_h)) Q_\star^{\ell}(y) + \delta_{\bar{y}_h}(y_h) c, \quad \forall (y_k, y_h) \in \mathcal{Y}_k \times \mathcal{Y}_h.$$

Note that $\hat{Q}^{\ell+1}$ does not sum up to one, but to $1 + c$, and therefore is not a probability distribution and does not lie in $\mathcal{M}_2^{\ell+1}$. In order to gain a probability distribution we then need to project this point to $\mathcal{M}_1^{\ell+1}$ via an *e*-projection. It is easy to see that this leads to a distribution in $\mathcal{M}_2^{\ell+1}$

$$P_0^{\ell+1}(y) = \hat{Q}^{\ell+1}(y_h | y_k) \tilde{P}(y_k),$$

for all $y \in \mathcal{Y}$. We then use this distribution as initial point for the *em*-algorithm between the spaces $\mathcal{M}_1^{\ell+1}$ and $\mathcal{M}_2^{\ell+1}$, as depicted in Figure 6.3.

In order to be able to compare this method to the traditional and following approaches we increase the size of the state space of the latent variables not only by one but by $2, 4$ and 8 states, such that we consider hidden spaces with 2, 4, 8 and 16 states. It remains to choose a constant c. We compare the outcome of the *em*-algorithm for 5 different random input distributions, in case of $|\mathcal{Y}_h| = 2$, and $c = 0.00001$, $c = 0.0001$, $c = 0.001$, $c = 0.01$ and $c = 0.1$ in Figure 6.4.

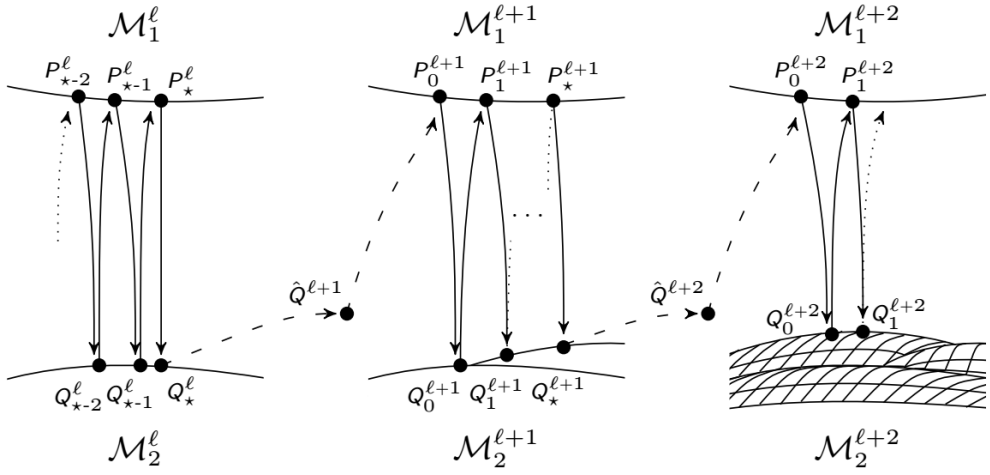


Figure 6.3: Sketch of the incrementally increasing *em*-algorithm using the natural method.

The resulting local minima are overall much more homogeneous in case of the natural method, such that the different minima form curves. The reason for this is that we only have 5 random input distributions for $\ell = 2$ and for each value of β. The initial distributions for $\ell > 2$ are completely determined by the result of the algorithm in the previous dimension and the choice of c. Hence we have fewer random influences compared to the traditional method. There we initialize the *em*-algorithm for each size of the state space with five random input distributions.

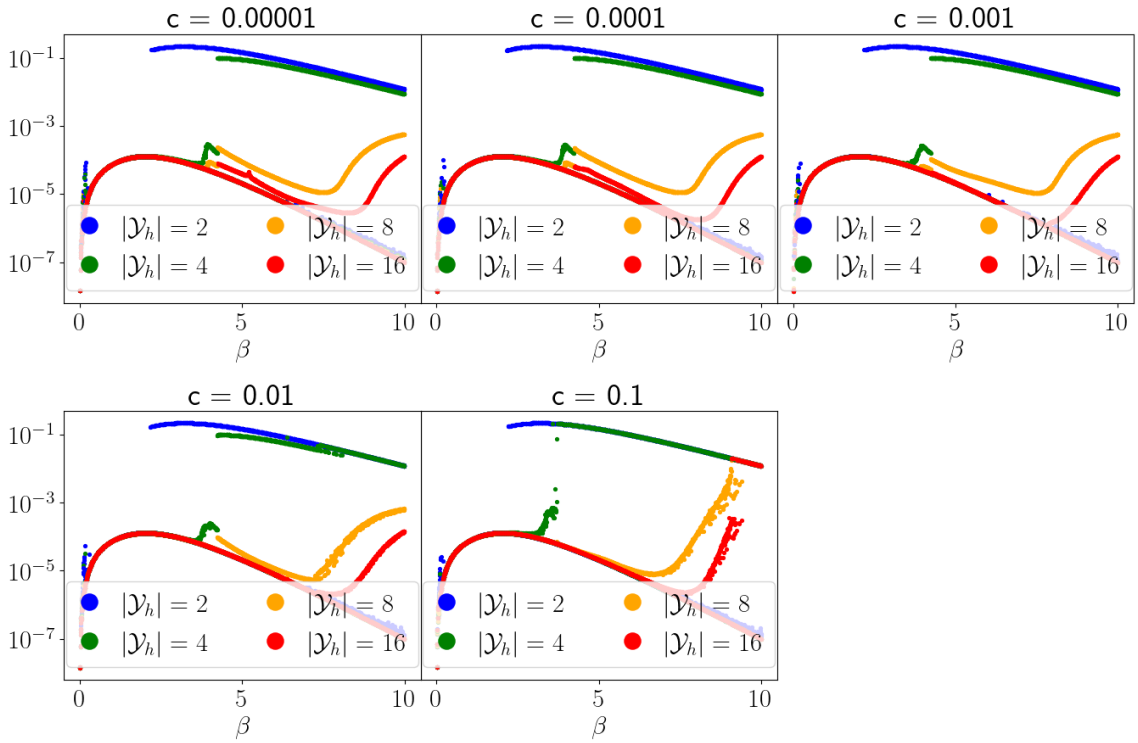


Figure 6.4: Results of the natural method for $c = 0.00001$, $c = 0.0001$, $c = 0.001$, $c = 0.01$ and $c = 0.1$.

Nonetheless, we observe that a larger constant c leads to more outliers on the curves, as we can see on the right in Figure 6.4 in case of $|\mathcal{Y}_h| = 8$ or 16. In that case there also exist local minima on the upper curve for large β and $|\mathcal{Y}_h| = 16$. Choosing a very small c, however , also does not lead the best outcome. Comparing the results for $c = 0.00001$ and $c = 0.001$ we see that the local minima in case of $|\mathcal{Y}_h| = 16$ are better for $c = 0.001$ overall. Hence, we choose $c = 0.001$ for the comparison with the other methods in Section 6.1.4.

Even though we are not able to guarantee a monotonic decrease of the local minima with increasing state space, we nonetheless observe that the results improve in general. The local minima for $\ell = 16$ are all smaller than $1.5 \cdot 10^{-4}$ for $c < 0.1$. Additionally, for $\beta > 7.5$ some values in the case of $\ell = 16$ lie on the lowest curve. We still observe some large local minima in the case of $\ell = 16$, but the curve consisting of these larger values starts to increase around $\beta = 8$, compared to $\beta = 7$ for the traditional method.

In the next section we introduce the "safe" method for which we can show that projecting to the larger space is guaranteed to lead to the same or an improved minimum.

6.1.2 Safe Method

For the definitions of the next two methods, namely the "safe" and "experimental" methods, we need to assume that we are able to divide the hidden variable Y_h into ℓ, $\ell \in \mathbb{N}$, binary variables such that one hidden state can be written as the vector $y_h^\ell = (y_{h,1}, \dots, y_{h,\ell})$. Then, after using the *em*-algorithm for a latent space with the size $|\mathcal{Y}_h^\ell| = 2^\ell$, we increase the size of the state space by one or more binary variables.

Note that here ℓ takes a slightly different role and with each increase of ℓ there is a significantly larger increase of the size of the state space compared to the previous method. In this case the state space grows from 2^ℓ states to $2^{\ell+1}$, instead of from ℓ to $\ell + 1$.

Here we define the new initial distribution $P_0^{\ell+1}$ in the larger space $\mathcal{M}_1^{2^{\ell+1}}$ using the result of the *em*-algorithm in $\mathcal{M}_1^{2^\ell}$ in the following way

$$P_0^{\ell+1}(y_k, y_h^{\ell+1}) = P_\star^\ell(y_k, y_h^\ell) \cdot R^{\ell+1}(y_{h,\ell+1} | y_h^\ell), \tag{6.1}$$

for all $(y_k, y_h^{\ell+1}) \in \mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}$ with a random distribution $R^{\ell+1} \in \mathcal{P}^\circ(\mathcal{Y}_{h^{\ell+1}})$. Then we use this $P_0^{\ell+1}$ as a starting point for the *em*-algorithm and project it to $\mathcal{M}_2^{2^{\ell+1}}$ using an *m*-projection, as depicted in Figure 6.5. This further decreases the KL-divergence.

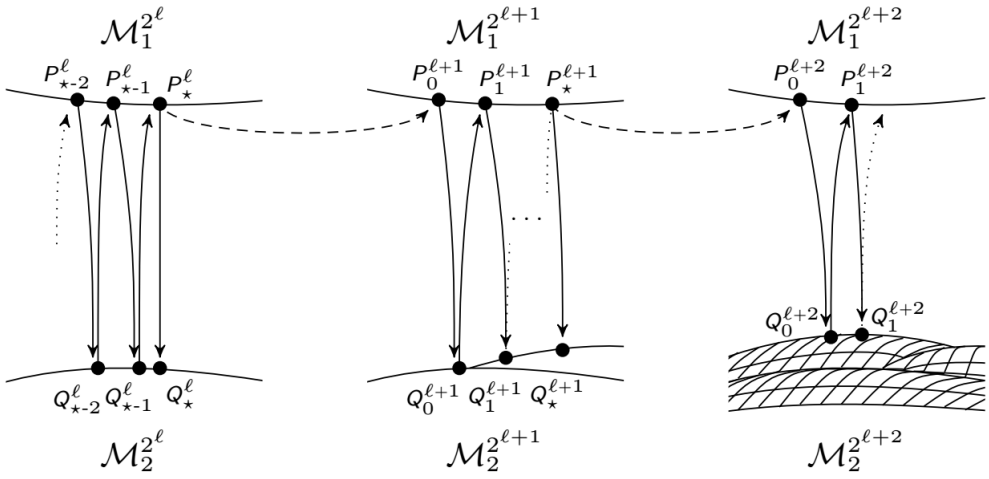


Figure 6.5: Sketch of the incrementally increasing *em*-algorithm for the safe or experimental method.

In Figure 6.6 we observe that this method leads to worse local minima compared to the natural method. The algorithm seems to stay fixed to a large local minimum found in the lower dimensional spaces. Nonetheless, we are able to show in the next proposition that using this approach guarantees to result in the same or an improved minimum in our setting.

Proposition 10. *Let $P_0^{\ell+1}$ be defined as in the equation (6.1). Additionally, we assume that $Q_\star^\ell(Y_k, Y_h^\ell) \otimes R^{\ell+1}(Y_{h,\ell+1}|Y_h^\ell) \in \mathcal{M}_2^{2^{\ell+1}}$. Then the following inequality holds*

$$D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell) \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}}(P_0^{\ell+1} \parallel Q_0^{\ell+1}),$$

where $Q_0^{\ell+1}$ is the m-projection of $P_0^{\ell+1}$ to $\mathcal{M}_2^{\ell+1}$.

The condition $Q_\star^\ell(Y_k, Y_h^\ell) \otimes R^{\ell+1}(Y_{h,\ell+1}|Y_h^\ell) \in \mathcal{M}_2^{\ell+1}$ is true when we make no additional assumption on the structure of the hidden variables. This is for example the case when $\mathcal{M}_2^{\ell+1}$ is given by a graphical model such that the hidden nodes are only connected among themselves and are the parents of some of the visible nodes. For the set $\mathcal{E}^m$, defined in (3.6) in the context of Φ_{CII}, we do not assume any structure on the hidden space and therefore the assumptions of Proposition 10 are met.

Proof of Proposition 10. We first extend the KL-divergence $D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell)$ to the larger space

$$\begin{aligned}
D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell) &= \sum_{y_k, y_h^\ell} P_\star^\ell(y_k, y_h^\ell) \log \frac{P_\star^\ell(y_k, y_h^\ell)}{Q_\star^\ell(y_k, y_h^\ell)} \\
&= \sum_{y_k, y_{h^\ell}} P_\star^\ell(y_k, y_{h^\ell}) \left(\sum_{y_{h,\ell+1}} R^{\ell+1}(y_{h,\ell+1}|y_h^\ell) \right) \log \frac{P_\star^\ell(y_k, y_{h^\ell})}{Q_\star^\ell(y_k, y_{h^\ell})} \\
&= \sum_{y_k, y_h^{\ell+1}} P_0^{\ell+1}(y_k, y_{h^{\ell+1}}) \log \frac{P_\star^\ell(y_k, y_h^\ell) \cdot R^{\ell+1}(y_{h,\ell+1}|y_h^\ell)}{Q_\star^\ell(y_k, y_h^\ell) \cdot R^{\ell+1}(y_{h,\ell+1}|y_h^\ell)} \\
&= \sum_{y_k, y_h^{\ell+1}} P_0^{\ell+1}(y_k, y_{h^{\ell+1}}) \log \frac{P_0^{\ell+1}(y_k, y_{h^{\ell+1}})}{Q_\star^\ell(y_k, y_h^\ell) \cdot R^{\ell+1}(y_{h,\ell+1}|y_h^\ell)}
\end{aligned}$$

Now the distribution $Q_\star^\ell(Y_k, Y_h^\ell) \otimes R^{\ell+1}(Y_{h,\ell+1}|Y_h^\ell) \in \mathcal{M}_2^{\ell+1}$ per assumption. Considering that $Q_0^{\ell+1}$ is the m-projection of $P_\star^\ell$ to $\mathcal{M}_2^{\ell+1}$, this results in the desired relationship

$$D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell) \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}}(P_0^{\ell+1} \parallel Q_0^{\ell+1}).$$

□

Note that this proposition is also true for $\ell = 1$. Now we are able connect the results from the *em*-algorithm for each size of the state space in the following way

$$\begin{aligned}
&D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_i^\ell \parallel Q_i^\ell) \geq \cdots \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell) \\
&\quad \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}}(P_0^{\ell+1} \parallel Q_0^{\ell+1}) \geq \cdots \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}}(P_\star^{\ell+1} \parallel Q_\star^{\ell+1}) \\
&\quad\quad \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+2}}(P_0^{\ell+2} \parallel Q_0^{\ell+2}) \geq \cdots \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+2}}(P_\star^{\ell+2} \parallel Q_\star^{\ell+2}) \geq \cdots .
\end{aligned}$$

Hence, by applying the safe method we gain a monotonically decreasing sequence of KL-divergences.

In particular, this proposition shows that the global minimum can not get larger when the size of the state space increases. However, the *em*-algorithm might result in the previously found minimum after increasing the state space and not converge to a better one, since this method introduces a random influence that only depends on the hidden variables. In Figure 6.6 we observe local minima larger than 10^{-2} for $|\mathcal{Y}_h| = 16$ and $\beta > 6$.

In order to address this caveat of the safe method we now introduce an approach in which the random distribution depends not only on the hidden but additionally on the known variables. We call this the "experimental" method.

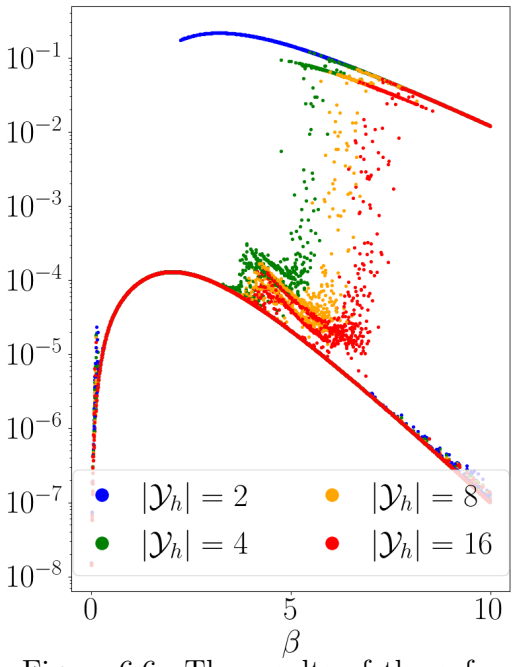


Figure 6.6: The results of the safe method.

6.1.3 Experimental Method

Finally, we introduce the "experimental" method. The approach of this method is closely related to the safe method, hence the sketch in Figure 6.5 also describes the steps taken in the experimental method.

Here the initial distribution $P_0^{\ell+1}$ is defined as follows

$$P_0^{\ell+1}(y_k, y_h^{\ell+1}) = P_\star^\ell(y_k, y_h^\ell) \cdot R^{\ell+1}(y_{h,\ell+1}|y_k, y_h^\ell), \tag{6.2}$$

for all $(y_k, y_h^{\ell+1}) \in \mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}$ with a random distribution $R^{\ell+1} \in \mathcal{P}(\mathcal{Y}_k \times \mathcal{Y}_{h^{\ell+1}})$. We are also able to prove a monotonically decreasing sequence, but here we need to use a stricter assumption compared to Proposition 10.

Proposition 11. *Let $P_0^{\ell+1}$ be defined as in (6.2). Furthermore, we assume that $Q_\star^\ell(Y_k, Y_h^\ell) \otimes R^{\ell+1}(Y_{h,\ell+1}|Y_k, Y_h^\ell) \in \mathcal{M}_2^{2^{\ell+1}}$, then the following inequality holds*

$$D_{\mathcal{Y}_k \times \mathcal{Y}_h^\ell}(P_\star^\ell \parallel Q_\star^\ell) \geq D_{\mathcal{Y}_k \times \mathcal{Y}_h^{\ell+1}}(P_0^{\ell+1} \parallel Q_0^{\ell+1})$$

where $Q_0^{\ell+1}$ is the m-projection of $P_0^{\ell+1}$ to $\mathcal{M}_2^{\ell+1}$.

This can be proven in the same manner as Proposition 10 in the case of the safe method. The stricter assumption that the distribution

$$Q_\star^\ell(Y_k, Y_h^\ell) \otimes R^{\ell+1}(Y_{h,\ell+1}|Y_k, Y_h^\ell)$$

lies in $\mathcal{M}_2^{\ell+1}$, however, is not met in the context of the measure Φ_{CII}.

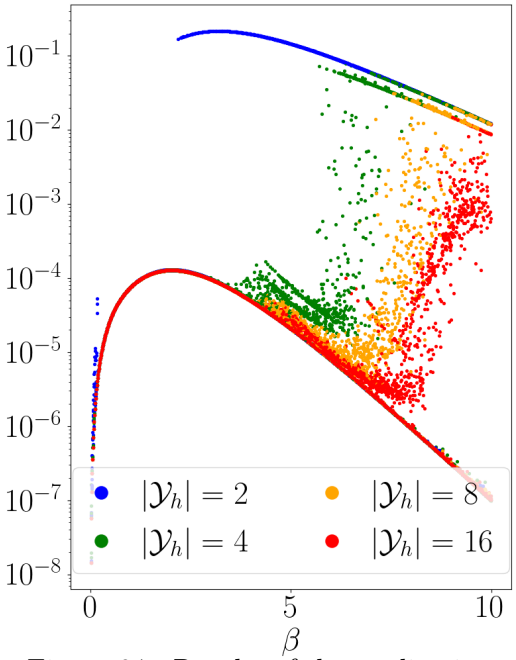


Figure 6.7: Results of the application of the experimental method.

Nonetheless, in Figure 6.7 we observe that the experimental method improves the efficiency of the safe method. Although there are still larger local minima for $\ell = 16$ and $\beta > 8$, there are fewer of those compared to the previous methods. Here the new initial distribution is perturbed more and therefore the algorithm can move away from the previous local minimum to a better one in the larger space more easily. We compare the different methods in more detail in the next section.

6.1.4 Comparison

We conclude this section with a comparison of the four different approaches in the setting of the example given by the Integrated Information measures Φ_{CII}. For each method we perform 5 runs with different random initial distributions and each approach performs exactly 1000 steps. The constant c used for the natural method is 0.001, because this yields the best results out of the five candidates that are compared in Figure 6.4.

The first row of Figure 6.8 depicts the local minima taken for each sized state space individually and the bottom row shows the minimum over all the runs and different sized state spaces. We connect the resulting local minima in the bottom row in order to increase the visibility of the sub-optimal outliers. Each column consists of the results for one of the methods.

As discussed in the context of the natural method, the traditional method uses 5 random input distributions for each sized state space and therefore the local minima have a greater variance compared to the other methods. Hence the minima in the top and bottom row do not lead to an approximation of a smooth curve for $\beta > 7$.

The natural method reduces the number of points that differ from the lower curve significantly, but we still see non-optimal local minima, especially for $\beta > 8$. The regularity of the points in case of the natural method is the result of choosing a fixed constant c. In this method the only random influence is the initial distribution for $|\mathcal{Y}| = 2$, whereas all the other methods have a random influence in each transition to a larger state space. Although this method works well, we are not able to prove that the minima do not increase with the size of the hidden space.

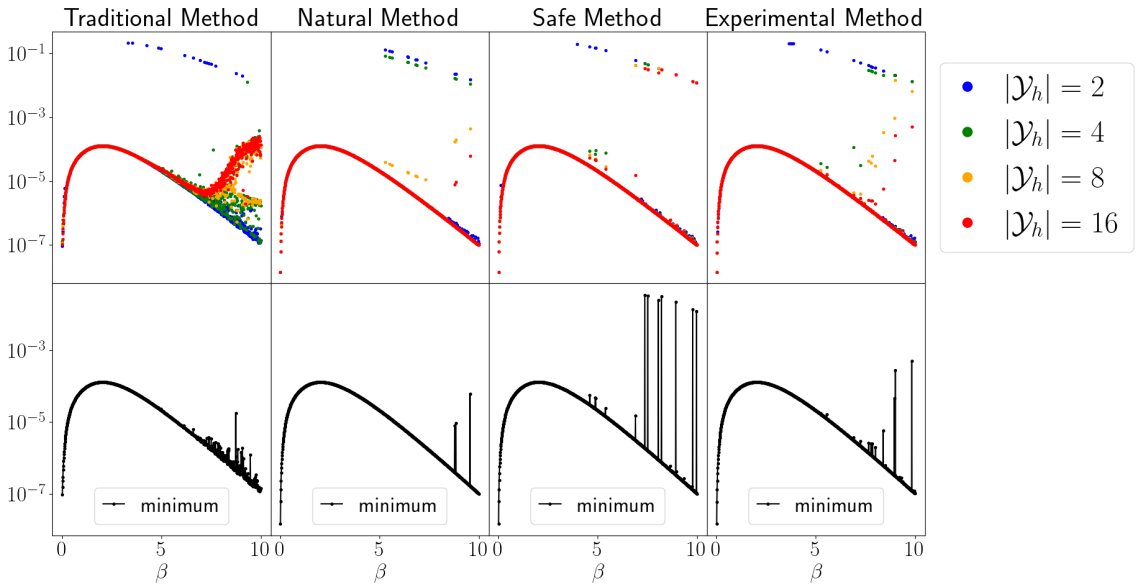


Figure 6.8: The local minima for each sized state space respectively on the top row and the local minimum over all the state spaces on the bottom.

For the safe method the random influence consists of $R^{\ell+1}(Y_{h,\ell+1}|Y_h^\ell)$ and here it is guaranteed that the local minimum does not increases with the size of the state space. We notice that there are still some large local minima for $\ell = 16$ in the top and therefore also in the bottom row. In these cases the random influence is not enough so that the algorithm stays fixed at a non-optimal minimum. Hence, we increase the random influence in the case of the experimental method. The results for this approach are depicted in the last column and it also exhibits non-optimal minima.

Overall, utilizing the local minima in a smaller space to find an initial distribution in the next larger space can lead to an improved result of the *em*-algorithm. From our experiments we can conclude that the experimental and natural method result in the least amount of large local minima, although we are not able to prove a monotonic decrease in the sequence of KL-divergences in these cases.

6.2 The Generalized Running Intersection Property

The classical iterative scaling algorithm, discussed in Section 2.5.2, can be applied to a type of exponential families that we refer to as "partition models". Hierarchical models are a subset of these models. In [Haberman74] the author defines the Running Intersection Property (RIP) for hierarchical models. This property guarantees that a model satisfying the RIP converges in one cycle to the MLE. In this section we define a new condition, called the Generalized Running Intersection Property (GRIP), which makes the analogues claim for a general partition model. The main part of these results are included in [Coons24] written in collaboration with Jane Ivy Coons and Michael Ruddy.

In order to understand the GRIP we first discuss hierarchical models and the RIP in more detail.

6.2.1 Hierarchical Models and the RIP

In Section 2.4.1 we define a hierarchical model as the exponential family corresponding to a simplicial complex. Here, we first discuss an additional representation of these models using matrices before we address the structure of a matrix associated with a hierarchical model.

The vector space $F = span\{f_0, \dots, f_{d-1}\}$ in the definition of the exponential family, Definition 4, can also be expressed as the rowspan of a matrix $A \in \mathbb{R}^{d\times|\mathcal{Y}|}$. Hence, every exponential family can be represented by such a matrix.

On the other hand, if we fix a matrix $A \in \mathbb{Z}^{d'\times|\mathcal{Y}|}$ such that the vector consisting of ones lies in the rowspan of A, then the set

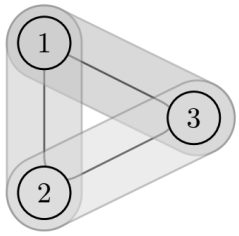
$$\mathcal{E}_A = \{P \in \mathcal{P}(\mathcal{Y})|\ \log(P) \in rowspan(A)\}$$

is called a log-linear model. Hierarchical models are log-linear models and we are able to create the corresponding matrix for every hierarchical model as follows. Let Γ be a simplicial complex with the facets $\{\gamma_0, \dots, \gamma_{d-1}\}$. The elements of Γ are associated with the vertices $v \in V$ that correspond to random variables with the state spaces $\mathcal{Y}_v$. Two vertices are connected by an edge if they lie together in a face in Γ. To simplify the notation we refer to the state space of the variables associated with $\gamma' \in facet(\Gamma)$ by $\mathcal{Y}_{\gamma'}$. Similarly, we denote by $\mathcal{Y}_\Gamma$ the combined state space of all the elements in V. Then the columns of the matrix $A_\Gamma \in \mathbb{Z}^{(\sum|\mathcal{Y}_{\gamma'}|)\times|\mathcal{Y}_\Gamma|}$ correspond to the elements of $\mathcal{Y}_\Gamma$, while the rows can be divided into d blocks, each connected to one facet. Let $y|_\gamma$ be the state of y restricted to the vertices in γ and let y_j^i be the jth state of the variables corresponding to the facet γ_i.

The elements of A in the jth row of the kth column of the ith block are defined in the following way

$$a^i_{j,k} = \begin{cases} 1, & \text{if } y_k|_{\gamma_i} = y^i_j \\ 0, & \text{otherwise.} \end{cases}$$

Example 4. Consider three binary variables and the simplicial complex Γ given by the facets $\{1,2\},\{2,3\}$ and $\{1,3\}$. The corresponding graphical representation and matrix are given in Figure 6.9. This hierarchical model was used in the context of the split systems of the synergistic measures Ψ_{Syn} in Section 4.5.3 and Ψ_{SynP} in Section 5.5.



$$A_\Gamma = \begin{array}{c} \begin{matrix} (0,0,0) & (0,0,1) & (0,1,0) & (0,1,1) & (1,0,0) & (1,0,1) & (1,1,0) & (1,1,1) \end{matrix} \\ \left(\begin{array}{cccccccc} \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} \\ \hdashline \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} \\ \hdashline \mathbf{1} & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \end{array}\right) \end{array} \begin{array}{l} (0,\,0,\,\cdot) \\ (0,\,1,\,\cdot) \\ (1,\,0,\,\cdot) \\ (1,\,1,\,\cdot) \\ (\cdot,\,0,\,0) \\ (\cdot,\,0,\,1) \\ (\cdot,\,1,\,0) \\ (\cdot,\,1,\,1) \\ (0,\,\cdot,\,0) \\ (0,\,\cdot,\,1) \\ (1,\,\cdot,\,0) \\ (1,\,\cdot,\,1) \end{array}$$

Figure 6.9: Example of a hierarchical model on the left and the corresponding matrix on the right.

Note that this construction leads to only one possible matrix corresponding to the hierarchical model. An exponential family can be expressed by various matrices with the same rowspan. However, the convergence of the iterative scaling algorithm depends heavily on the properties of the chosen matrix. The following example emphasizes this.

Example 5. Here we consider the two matrices A and $\tilde{A}$.

$$A = \left(\begin{array}{ccc} \mathbf{1} & \mathbf{1} & 0 \\ \hdashline 0 & 0 & \mathbf{1} \\ \hdashline \mathbf{1} & 0 & 0 \\ \hdashline 0 & \mathbf{1} & \mathbf{1} \end{array}\right) \qquad \tilde{A} = \begin{pmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}$$

Both matrices have rowspan over $\mathbb{R}$ equal to $\mathbb{R}^3$, hence $\mathcal{E}_A = \mathcal{E}_{\tilde{A}}$.

Although they represent the same model, the convergence of the IPS algorithm is heavily influenced by the chosen representation. Let $d = (d_1, d_2, d_3)$ be the normalized data vector. Using matrix $\tilde{A}$ the iterative scaling algorithm converges in one step to the MLE, given by $P^\star = d$. Here we denote the distributions in form of vectors $P = (P(y_1), P(y_2), P(y_3))$.

$$P^0 = \left(\frac{1}{2}(d_1+d_2),\, \frac{1}{2}(d_1+d_2), d_3\right)$$

$$P^1 = \left(d_1,\, \frac{1}{2}(d_1+d_2)\frac{(d_2+d_3)}{\frac{1}{2}(d_1+d_2)+d_3}, d_3\frac{(d_2+d_3)}{\frac{1}{2}(d_1+d_2)+d_3}\right)$$

In general the different projections have the following form:

$$P^{\ell} = \left(d_1, \quad P^{\ell-1}(y_2)\frac{(d_2+d_3)}{P^{\ell-1}(y_2)+d_3}, d_3\frac{(d_2+d_3)}{P^{\ell-1}(y_2)+d_3} \right), \quad \ell \text{ odd} \tag{6.3}$$

$$P^{\ell} = \left(d_1\frac{(d_1+d_2)}{d_1+P^{\ell-1}(y_2)}, P^{\ell-1}(y_2)\frac{(d_1+d_2)}{d_1+P^{\ell-1}(y_2)}, d_3 \right), \quad \ell \text{ even} \tag{6.4}$$

Suppose that there exists an index ℓ such that the second entry of P^{ℓ} is exactly d_2 in (6.3) as well as (6.4), then we can deduce that $d_2 = P^{\ell-1}(y_2)$. Hence, the IPS algorithm can only result in the exact result if $d_1 = d_2$. Note that while the model is symmetric, the application of the IPS algorithm is not. Hence, if we would reorder the partitions, then the exact convergence would require $d_2 = d_3$.

In an evaluation with 20 000 random input distributions d, the arithmetic mean of the iteration steps taken till a step size smaller than 10^{-8} was reached results in approx 113 steps with a minimum value of 8 and a maximum value of 287 478. Recall that in case of $\tilde{A}$ the necessary number of iteration steps is exactly 1.

The previous example shows the significant impact that the matrix representation can have on the convergence of the iterative scaling algorithm. Hence, we now introduce a special ordering to the facets, called the running intersection property.

Definition 21 (Running Intersection Property)**.** Let $O = \{F_1, \ldots, F_s\}$ be an ordering of the facets of a simplicial complex Γ. Then this ordering satisfies the running intersection property (RIP), if for each $r \in \{1, \ldots, s\}$ there exists a k_r such that

$$\left(\bigcup_{k=1}^{r} F_k \right) \cap F_{r+1} = F_{k_r} \cap F_{r+1}.$$

One class of simplicial complexes with an ordering that satisfies the RIP is called decomposable.

Definition 22 (Decomposable Simplicial Complex)**.** A simplicial complex Γ is called decomposable if it has only one facet or if it is the union of two decomposable simplicial complexes Γ_1, Γ_2 such that there exist facets $F_1 \in \Gamma_1$ and $F_2 \in \Gamma_2$ with

$$\bigcup_{F \in facet(\Gamma_1)} F \cap \bigcup_{\tilde{F} \in facet(\Gamma_2)} \tilde{F} = F_1 \cap F_2.$$

The connection between decomposable simplicial complexes and the running intersection property, as stated in the Lemma below, is proven for example as Lemma 5.10 in [Haberman74].

Lemma 1. *Let Γ be a decomposable simplicial complex. Then there exists an ordering of the facets that satisfies the RIP.*

The following Theorem draws a connection between decomposability of a simplicial complex and one-cycle convergence of the iterative scaling algorithm applied to the corresponding hierarchical model. One-cycle convergence can be achieved by requiring that the matrix associated with the hierarchical model is ordered such that the facets corresponding to the partition matrices satisfy the RIP.

Theorem 6.2.1 (Theorem 5.3 in [Haberman74]). *Suppose the simplicial complex Γ is decomposable. Let $A = A^{1,\dots,k}$ be the matrix corresponding to an ordering of the facets of Γ that satisfies the RIP. Then the iterative scaling algorithm applied to A yields the MLE in one cycle.*

There exist stronger formulations of this property, which guarantee one-cycle convergence for every ordering of a simplicial complex. These are then called totally decomposable models and this is proven as Theorem 5.4 in [Haberman74]. An additional condition for strong decomposability is proven in [Vomlel99].

In the following section, we define a generalized version of the running intersection property that can be applied to a broader range of partition models. We show that the matrices that satisfy the generalized running intersection property exhibit one-cycle convergence under the iterative scaling algorithm.

6.2.2 Partition Models and the GRIP

In this section we discuss the more general set of exponential families called partition models and a criterion for one-cycle convergence. First we define the type of matrices that represent partition models.

Definition 23 (Multipartition Matrix). A matrix $A \in \{0,1\}^{n\times m}$ with $n, m \in \mathbb{N}$ is a multipartition matrix if one can partition the rows of A into submatrices $A^1, \dots, A^d$ such that in each submatrix the entries of every column sum to 1. The matrices $A^1, \dots, A^d$ are called the partition matrices of A.

The log-linear model corresponding to a multipartition matrix is called partition model. Note that this definition of a partition model is equivalent to the one given in Section 2.5.2, because for each set of partitions $\Pi^{1,\dots,\nu}$, there exists a multipartition matrix, A_Π, representing the corresponding model, $\mathcal{E}(\Pi^{1,\dots,\nu})$, in the following way. Each partition $\Pi^i(\mathcal{Y})$ corresponds to a block in the matrix A_Π and each element in the jth row and kth column of the ith block results in

$$a^i_{j,k} = \begin{cases} 1, & \text{if } y_k \in \Pi^i_j(\mathcal{Y}) \\ 0, & \text{otherwise,} \end{cases}$$

where $\Pi^i_j(\mathcal{Y})$ is the jth set of the ith partition.

Similarly, it is possible to find the set of partitions corresponding to a multipartition matrix by using each block as a partition and each row as a set in this partition. Since each column sums to one for each block, this leads to a set of partitions. The following example highlights this connection between these two concepts.

Example 6. Consider the following example with three partitions

$$\begin{aligned}\Pi^{1,2,3} = \Big\{ &\Big\{\{y_1,y_2,y_3,y_4,y_5,y_6,y_7\},\{y_8,y_9,y_{10},y_{11},y_{12},y_{13},y_{14}\}\Big\},\\ &\Big\{\{y_1,y_2,y_3,y_8,y_9,y_{10}\},\{y_4,y_5,y_6,y_7,y_{11},y_{12},y_{13},y_{14}\}\Big\},\\ &\Big\{\{y_1,y_8\},\{y_2,y_9\},\{y_3,y_{10}\},\{y_4,y_6,y_{11},y_{13}\},\{y_5,y_7,y_{12},y_{14}\}\Big\}\Big\}.\end{aligned}$$

Following the construction described above results in the matrix A_Π depicted in Figure 6.10, where each partition corresponds to a block in the matrix A_Π.

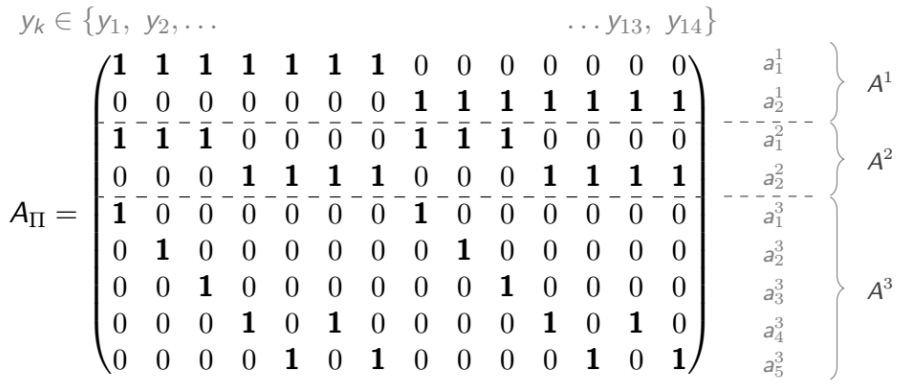


Figure 6.10: Example of a multipartition matrix.

This matrix has repeated columns, namely column 4 and 6, 5 and 7, 11 and 13, as well as 12 and 14 are equal. The impact of repeated columns is discussed in more detail in Proposition 3.25 and Example 3.26 in [Coons24]. There we show that removing all repeated columns from a matrix satisfying the GRIP, which we introduce in Definition 29, results in another multipartition matrix fulfilling the GRIP, while the reverse is not true. Repeating a column can destroy the structure necessary for the one-cycle convergence.

The matrices A_Γ that correspond to hierarchical models are multipartition models and therefore hierarchical models are a subset of partition models.

The author of [Rauh13] names the log-linear model associated to only one partition matrix "partition model". He studies these models as an example of a family of log-linear models for which there exists a low constant that bounds the KL-divergence between an arbitrary point in the simplex and these models. Thereby he proves additional properties of the models defined by one partition matrix.

In order to be able to define the generalized running intersection property we first need to introduce additional concepts regarding the structure of a multipartition model. The first property is called "floret condition". Here the naming is deliberately suggestive of the terminology for staged trees. In the Section 4.2 of [Coons24] the connection between the GRIP and staged trees is highlighted in more detail and it is shown that the tree associated to the partition matrix satisfying the GRIP is a stratified and balanced staged tree. An introduction to staged trees can be found in, for example, [Collazo18].

Definition 24 (Floret Condition)**.** Let B and C be two partition matrices with the same number of columns and with rows a_j^B and $a_{\tilde{j}}^C$. The rows a_j^B and $a_{\tilde{j}}^C$ are called connected if there is at least one column index k such that the kth index of a_j^B and $a_{\tilde{j}}^C$ are both 1.

The matrices B and C satisfy the floret condition if for every two rows of B, a_j^B and $a_{j'}^B$, the sets of rows of C that are connected to a_j^B and $a_{j'}^B$ are disjoint or equal. The set of rows of B connected to a fixed row of C is called a floret of B and the set of rows of C connected to a fixed row of B is a floret of C.

Example 7. The multipartition matrix on the right does not satisfy the floret condition, since the first row a_1^1 is connected to both rows in the second partition, while the second row a_1^2 is only connected to a_1^2.

$$\left(\begin{array}{ccccccc} \mathbf{1} & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ \hline \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{array}\right)$$

Next, we define notation regarding the repeated columns in a multipartition matrix.

Definition 25 (Column Weight)**.** For a multipartition matrix A the column weight of the k-th column is the number of times the column is repeated. We denote this by $c_k \in \mathbb{Z}_+$.

Let $I(a_j^i)$ be the set of all column indices k for the row a_j^i such that its kth entry is equal to 1. This can be identified with the jth set in the ith partition of the set of partitions corresponding to the multipartition matrix A. In Example 6 the index sets $I(a_1^2) = \{1, 2, 3, 8, 9, 10\}$ and $I(a_1^3) = \{1, 8\}$. Additionally, let $\mathcal{S}(i, k)$ be the function that maps the ith partition and the kth column to the row such that $k \in I(a^i_{\mathcal{S}(i,k)})$. Hence $a^i_{\mathcal{S}(i,k)}$ has a 1 in the kth entry.

The following definition is motivated by the way column weights behave under steps of the iterative scaling algorithm.

Definition 26 (Well-Connected and Connection Ratio)**.** For a multipartition matrix A the c_k^ℓ is defined as the k-th column weight for the matrix obtained by only considering the first ℓ partitions of A. Then A is well-connected if for any row vector a_j^ℓ, with $\ell > 1$, the equality $(c_k^\ell / c_k^{\ell-1}) = (c_{k'}^\ell / c_{k'}^{\ell-1})$ holds for all $k, k' \in I(a_j^\ell)$. We call this quantity the connection ratio for a_j^ℓ and denote it by C_j^ℓ with the convention that $C_j^1 = |I_j^1|$.

The matrix A_Π in Example 6 is well-connected. The connection ratios for the rows are in this case the following. $a_1^1 = a_2^1 = 7$, $a_1^2 = 3/7$, $a_2^2 = 4/7$, $a_1^3 = a_2^3 = a_3^3 = 1/3$ and $a_4^3 = a_5^3 = 2/4$.

It remains to define two operations on the partition matrices of a multipartition model that play a similar role as joining and intersecting the facets of the simplicial complex in the RIP.

Definition 27 (Union of two Partition Matrices)**.** The union of two partition matrices $B = A^1 \Cup A^2$ is defined as a partition matrix in which there exist for every row b of B a row a_j^1 of A^1 and a row $a^2_{\tilde{j}}$ of A^2 such that $I(b) = I(a_j^1) \cap I(a^2_{\tilde{j}})$.

The next example illustrates this concept.

Example 8. Let A^1 and A^2 be the partition matrices from Example 6, then the union of these matrices results in

$$A^1 \Cup A^2 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix}$$

The first two rows can be gained by intersecting the index sets of a_1^1 with a_1^2 and a_2^2, while the index sets of the third and fourth row result from the intersections between the index sets of a_2^1 with a_1^2 and a_2^2 respectively.

Definition 28 (Intersection of two Partition Matrices)**.** Suppose that A^1 and A^2 satisfy the floret condition and let $F_1^2, \ldots, F_{d'}^2$ be the distinct florets of A^2 with respect to A^1. We define the matrix $B = A^1 \Cap A^2$ to be the matrix where each row corresponds to one of these florets in the following way. The index set of a row $b_{j'}$ of B is the union over the index sets of the rows in the corresponding floret $\mathcal{F}_{j'}^2$. Therefore the index set of a row $b_{j'}$ of B results in

$$I(b_{j'}) = \bigcup_{a^2_{\tilde{j}} \in \mathcal{F}^2_{j'}} I(a^2_{\tilde{j}}).$$

In other words, the columns of $A^1 \Cap A^2$ are indicator vectors for the florets that each column's non-zero rows belong to. Note that the intersection is only defined when both matrices satisfy the floret condition.

Example 9. Let A^1 and A^2 be the partition matrices from Example 6, then the partition matrix intersection results in

$$A^2 \Cap A^3 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix}$$

With these definitions we are now able to define the GRIP.

Definition 29 (GRIP)**.** Let $A^{1,\dots,d}$ be a multipartition matrix. For each $\ell \in \{1,\dots,d\}$, let B_ℓ denote $\Cup_{i=1}^{\ell} A^i$. Then $A^{1,\dots,d}$ satisfies the generalized running intersection property, or GRIP if for each $1 \leq \ell \leq k-1$,

1. the matrix $B_\ell A^{\ell+1}$ is well-connected,
2. $B_\ell A^{\ell+1}$ satisfies the floret condition, and
3. the rows of $B_\ell \Cap A^{\ell+1}$ lie in the rowspan of $A^{1,\dots,\ell}$.

Note that the third condition in the definition above resembles the running intersection property. In fact, translating the conditions of the RIP from facets to their corresponding matrices leads exactly to the requirement above. The first and second condition are additionally necessary here, because we require a certain structure that hierarchical models already satisfy per definition. This leads us to the main result of this chapter.

Theorem 6.2.2. *If A is a multipartition matrix with k partitions that satisfies the GRIP given in Definition 29, then the iterative scaling algorithm results in the MLE after one cycle.*

In [Coons24] we prove the theorem above by first deriving a general formula for the MLE of models with matrices that satisfy the GRIP. Using this, we are able to prove the one-cycle convergence. There we apply methods from algebraic geometry, such as the toric fiber product. Here we instead recreate the proof strategy applied by Haberman in [Haberman74] for the one-cycle convergence in the case of the RIP. This method is more technical, but it shows the strong resemblance between the RIP and GRIP and requires less background knowledge in algebraic geometry.

Before we prove this result we show in the next example how the iterative scaling algorithm builds the MLE step vise for matrices that satisfy the GRIP.

In order to simplify the notation we denote

$$a^i_j(d) = \sum_k a^i_{j,k} d_k.$$

Example 10. Here we apply the iterative scaling algorithm to the matrix A_Π from Example 6. This matrix satisfies the GRIP.

Let $d = (d_1,\dots,d_{14})$ be the normalized data vector. Projecting to the first partition of A leads to:

$$P^0(y_1) = \cdots = P^0(y_7) = \frac{1}{7}\, a^1_1(d) \quad \text{and} \quad P^0(y_8) = \cdots = P^0(y_{14}) = \frac{1}{7}\, a^1_2(d).$$

The second step of the algorithm results in four different types of indices. These are given by the different rows in the matrix $A^1 \Cup A^2$, indicated by the different dashed and dotted lines below the matrix.

$$A^1 \Cap A^2 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix} \quad a_1^{1 \Cap 2}$$

$$A^1 \Cup A^2 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix} \begin{matrix} a_1^{1 \Cup 2} \\ a_2^{1 \Cup 2} \\ a_3^{1 \Cup 2} \\ a_4^{1 \Cup 2} \end{matrix}$$

Note that since $A^1 \Cup A^2$ consists only of the ones vector and d is normalized, we have $a_1^{1 \Cap 2}(d) = 1$. Therefore the second projection simplifies in the following way:

$$\begin{aligned}
P^1(y_1) = P^1(y_2) = P^1(y_3) &= \frac{1}{C_1^1} a_1^1(d) \frac{a_1^2(d)}{C_1^2 \, a_1^{1 \Cap 2}(d)} = \frac{1}{3} a_1^1(d) \cdot a_1^2(d) \\
P^1(y_4) = P^1(y_5) = P^1(y_6) = P^1(y_7) &= \frac{1}{C_1^1} a_1^1(d) \frac{a_2^2(d)}{C_2^2 \, a_1^{1 \Cap 2}(d)} = \frac{1}{4} a_1^1(d) \cdot a_2^2(d) \\
P^1(y_8) = P^1(y_9) = P^1(y_{10}) &= \frac{1}{C_2^1} a_2^1(d) \frac{a_1^2(d)}{C_1^2 \, a_1^{1 \Cap 2}(d)} = \frac{1}{3} a_2^1(d) \cdot a_1^2(d) \\
P^1(y_{11}) = P^1(y_{12}) = P^1(y_{13}) = P^1(y_{14}) &= \frac{1}{C_2^1} a_2^1(d) \frac{a_2^2(d)}{C_2^2 \, a_1^{1 \Cap 2}(d)} = \frac{1}{4} a_2^1(d) \cdot a_2^2(d)
\end{aligned}$$

For the last projection we only demonstrate four indices as an example.

$$(A^1 \Cup A^2) \Cap A^3 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix} \begin{matrix} \alpha_1^{(1 \Cup 2) \Cap 3} \\ \alpha_2^{(1 \Cup 2) \Cap 3} \end{matrix}$$

In this case $A^2 = (A^1 \Cup A^2) \Cap A^3$, hence $\alpha_i^2(d) = \alpha_i^{(1 \Cup 2) \Cap 3}(d)$. In order to visualize the structure created by the IPS we also display the general formula of p_j^2 without simplifications

$$\begin{aligned}
P^2(y_1) &= \frac{1}{C_1^1} a_1^1(d) \frac{a_1^2(d)}{C_1^2 \, a_1^{1 \Cap 2}(d)} \frac{a_1^3(d)}{C_1^3 a_1^{(1 \Cup 2) \Cap 3}(d)} \\
&= \frac{1}{7} a_1^1(d) \frac{a_1^2(d)}{\frac{3}{7} a_1^{1 \Cap 2}(d)} \frac{a_1^3(d)}{\frac{1}{3} a_1^{(1 \Cup 2) \Cap 3}(d)} = a_1^1(d) \cdot a_1^3(d) \\
P^2(y_4) &= \frac{1}{C_1^1} a_1^1(d) \frac{a_2^2(d)}{C_2^2 \, a_1^{1 \Cap 2}(d)} \frac{a_4^3(d)}{C_4^3 a_2^{(1 \Cup 2) \Cap 3}(d)} \\
&= \frac{1}{7} a_1^1(d) \frac{a_2^2(d)}{\frac{4}{7} a_1^{1 \Cap 2}(d)} \frac{a_4^3}{\frac{2}{4} a_2^{(1 \Cup 2) \Cap 3}(d)} = \frac{1}{2} a_1^1(d) \cdot a_4^3(d) \\
P^2(y_8) &= \frac{1}{C_2^1} a_2^1(d) \frac{a_1^2(d)}{C_1^2 \, a_1^{1 \Cap 2}(d)} \frac{a_1^3(d)}{C_1^3 a_1^{(1 \Cup 2) \Cap 3}(d)} \\
&= \frac{1}{7} a_2^1(d) \frac{a_1^2(d)}{\frac{3}{7} a_1^{1 \Cap 2}(d)} \frac{a_1^3(d)}{\frac{1}{3} a_1^{(1 \Cup 2) \Cap 3}(d)} = a_2^1(d) \cdot a_1^3(d) \\
P^2(y_{11}) &= \frac{1}{C_2^1} a_2^1(d) \frac{a_2^2(d)}{C_2^2 \, a_1^{1 \Cap 2}(d)} \frac{a_4^3(d)}{C_4^3 a_2^{(1 \Cup 2) \Cap 3}(d)} \\
&= \frac{1}{7} a_2^1(d) \frac{a_2^2(d)}{\frac{4}{7} a_1^{1 \Cap 2}(d)} \frac{a_4^3}{\frac{2}{4} a_2^{(1 \Cup 2) \Cap 3}(d)} = \frac{1}{2} a_1^1(d) \cdot a_4^3(d)
\end{aligned}$$

For the indices 4 and 11 there exists two identical columns, respectively. Hence the connection ratios do not add up to one, but to $\frac{1}{2}$.

In order to prove the Theorem 6.2.2, we first need to introduce the following monomial notation for elements of $\mathcal{E}_A$, as discussed in more detail in for example Section 6.2 of [Sullivant18] or Section 1.2 of [Drton09a]. Let A be a partition matrix with n rows and m columns and denote the element in the jth row and kth column by $a_{j,k}$. For each element $P \in \mathcal{E}_A$ there exist $t = (t_1, \dots, t_n) \in \mathbb{R}^n$, such that

$$P = \left(\prod_{j=1}^{n} t_j^{a_{j,1}}, \prod_{j=1}^{n} t_j^{a_{j,2}}, \dots, \prod_{j=1}^{n} t_j^{a_{j,m}} \right).$$

Now we are able to prove the first step towards Theorem 6.2.2. First, we show in the next proposition that given some regularity assumptions we are able to write every distribution in $\mathcal{E}_A$ as a product of marginal distributions divided by their intersection.

Proposition 12. *Let A be a matrix consisting of the partitions $A^1, \dots, A^d$ such that the matrix consisting of $B = \biguplus_{i=1}^{\ell} A^i$ and $C = \biguplus_{i=\ell+1}^{d} A^i$ satisfies the floret condition and is well connected. We denote the rows of B by a^B, the rows of C by a^C and the rows of*

$$D := B \Cap C$$

by a^D. Then every $P \in \mathcal{M}_A$ can be written as

$$P(y_k) = \frac{\left(a^B_{\mathcal{S}(k)}(P)\right)\left(a^C_{\mathcal{S}(k)}(P)\right)}{c_k^d\left(a^D_{\mathcal{S}(k)}(P)\right)},$$

for all $y_k \in \mathcal{Y}$ and where $a^B_{\mathcal{S}(k)}$ denotes the row of B with a 1 at column k, $k \in I(a^B_{\mathcal{S}(k)})$.

Proof. Let n_i be the number of rows of the partition matrix A^i. Then every $P(y_k) \in \mathcal{E}_A$ can be written as

$$P(y_k) = \prod_{i=1}^{d} \prod_{j=1}^{n_i} t_j^{a^i_{j,k}}, \quad \forall y_k \in \mathcal{Y}.$$

Since every matrix $A^1, \dots, A^d$ is a partition matrix, there only exists one $t_j^{a^i_{j,k}} \neq 1$ for each $i \in \{1, \dots, d\}$, hence we use the shorter notation $t_{\mathcal{S}(i,k)}$

$$P(y_k) = \prod_{i=1}^{d} t_{\mathcal{S}(i,k)}, \quad \forall y_k \in \mathcal{Y}.$$

Therefore, we are able to write

$$P(y_k) = \prod_{i=1}^{d} t_{\mathcal{S}(i,k)} = \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k)} \right) \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k)} \right), \quad \forall y_k \in \mathcal{Y}.$$

Now the elements in the marginal distribution w.r.t. B can be written as follows:

$$a^B_{\mathcal{S}(k)}(P) = \sum_{k' \in I(a^B_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')} \right) \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')} \right).$$

The elements in B are created by intersecting the index sets of the elements in $A^1, \dots, A^\ell$, therefore the first product is the same for every index $k' \in I(a^B_{\mathcal{S}(k)})$.

$$a^B_{\mathcal{S}(k)}(P) = \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k)}\right) \sum_{k' \in I(a^B_{\mathcal{S}(k)})} \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')}\right).$$

Now we are able to rewrite the sum to sum over the indices of the rows in C, that are connected to $a^B_{\mathcal{S}(k)}$, multiplied by the column weight c^d_k. Let $F^C(a^B_{\mathcal{S}(k)})$ be the set of rows in C connected to $a^B_{\mathcal{S}(k)}$. Let $J^C(a^B_{\mathcal{S}(k)})$ be the set consisting of exactly one index in $I(a^C) \cap I(a^B_{\mathcal{S}(k)})$ for each $a^C \in F^C(a^B_{\mathcal{S}(k)})$. Note that c^ℓ_k is the column weight in the matrix B. Since B consists of only one partition, the column weight $c^\ell_{k'}$ is the same value for every $k' \in J^C(a^B_{\mathcal{S}(k)})$.

$$\begin{aligned} a^B_{\mathcal{S}(k)}(P) &= \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k)}\right) \sum_{k' \in J^C(a^B_{\mathcal{S}(k)})} c^d_{k'} \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')}\right) \\ &= \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k)}\right) c^\ell_k \sum_{k' \in J^C(a^B_{\mathcal{S}(k)})} \frac{c^d_{k'}}{c^\ell_{k'}} \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')}\right) \end{aligned} \tag{6.5}$$

Since the matrices B and C are well connected and satisfy the floret condition, the sum in (6.5) is the same for every row in B belonging to the same floret. Now we do the same operations for $a^C_{\mathcal{S}(k)}$

$$\begin{aligned} a^C_{\mathcal{S}(k)}(P) &= \sum_{k' \in I(a^C_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')}\right) \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')}\right) \\ &= \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k)}\right) \sum_{k' \in I(a^C_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')}\right) \\ &= \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k)}\right) \sum_{k' \in J^B(a^C_{\mathcal{S}(k)})} c^d_{k'} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')}\right) \\ &= \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k)}\right) \frac{c^d_k}{c^\ell_k} \sum_{k' \in J^B(a^C_{\mathcal{S}(k)})} c^\ell_{k'} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')}\right). \end{aligned}$$

Let a^D be a row in D. As discussed after Definition 28, we can think of the rows of D as indicator vectors for the florets. Hence this leads to the following formulas

$$\begin{aligned} a^D_{\mathcal{S}(k)}(P) &= \sum_{k' \in I(a^D_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')}\right) \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')}\right) \\ &= \sum_{k' \in J^B(a^D_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')}\right) \sum_{k'' \in I(a^B_{\mathcal{S}(k')})} \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k'')}\right) \end{aligned}$$

and therefore to

$$\begin{aligned} a^D_{\mathcal{S}(k)}(P) &= \sum_{k' \in J^B(a^D_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')} \right) \sum_{k'' \in J^C(a^B_{\mathcal{S}(k')})} \left(\prod_{i=\ell+1}^{d} c^d_{k''} t_{\mathcal{S}(i,k'')} \right) \\ &= \sum_{k' \in J^B(a^D_{\mathcal{S}(k)})} \left(\prod_{i=1}^{\ell} c^\ell_{k'} t_{\mathcal{S}(i,k')} \right) \sum_{k'' \in J^C(a^B_{\mathcal{S}(k')})} \left(\prod_{i=\ell+1}^{d} \frac{c^d_{k''}}{c^\ell_{k''}} t_{\mathcal{S}(i,k'')} \right). \end{aligned}$$

Because B and C satisfy the floret condition and are well connected, the second sum is the same for every element in $J^B(a^D_{\mathcal{S}(k)})$, which leads to

$$a^D_{\mathcal{S}(k)}(P) = \left(\sum_{k' \in J^B(a^C_{\mathcal{S}(k)})} \prod_{i=1}^{\ell} c^\ell_{k'} t_{\mathcal{S}(i,k')} \right) \left(\sum_{k' \in J^C(a^B_{\mathcal{S}(k)})} \prod_{i=\ell+1}^{d} \frac{c^d_{k'}}{c^\ell_{k'}} t_{\mathcal{S}(i,k')} \right).$$

In conclusion we have

$$\begin{aligned} \frac{\left(a^B_{\mathcal{S}(k)}(P)\right)\left(a^C_{\mathcal{S}(k)}(P)\right)}{\left(a^D_{\mathcal{S}(k)}(P)\right)} &= \frac{\left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k)} \right) c^\ell_k \sum_{k' \in J^C(a^B_{\mathcal{S}(k)})} \frac{c^d_{k'}}{c^\ell_{k'}} \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')} \right)}{\left(\sum_{k' \in J^B(a^C_{\mathcal{S}(k)})} \prod_{i=1}^{\ell} c^\ell_{k'} t_{\mathcal{S}(i,k')} \right) \left(\sum_{k' \in J^C(a^B_{\mathcal{S}(k)})} \prod_{i=\ell+1}^{d} \frac{c^d_{k'}}{c^\ell_{k'}} t_{\mathcal{S}(i,k')} \right)} \\ &\quad \cdot \left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k)} \right) \frac{c^d_k}{c^\ell_k} \sum_{k' \in J^B(a^C_{\mathcal{S}(k)})} c^\ell_{k'} \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k')} \right) \\ &= c^d_k P(y_k) \end{aligned}$$

for all $y_k \in \mathcal{Y}$. □

The next step is to show that the MLE can be written in the same way as a combination of the MLEs corresponding to the matrices B, C and D.

Proposition 13. *Let A be a multipartition matrix consisting of the partitions $A^1, \ldots, A^d$ such that the matrix consisting of $B = \biguplus_{i=1}^{\ell} A^i$ and $C = \biguplus_{i=\ell+1}^{d} A^i$ satisfies the floret condition and is well connected and such that the rows of $D = B \Cap C$ lie in the rowspan of $A^{1,\ldots,\ell}$. Now let $P^\star \in \mathcal{E}_A$ be the MLE of a positive distribution P, then we are able to write*

$$P^\star(y_k) = \frac{\left(\frac{a^B_{\mathcal{S}(k)}(P)}{c^\ell_k} \right)^\star \left(\frac{a^C_{\mathcal{S}(k)}(P)}{c^C_k} \right)^\star}{\frac{c^d_k}{c^\ell_k c^C_k} a^D_{\mathcal{S}(k)}(P)},$$

for all $y_k \in \mathcal{Y}$ and where c^C_k is the connection number of the k-th column in the matrix C and where $\left(\frac{a^B_{\mathcal{S}(k)}(P)}{c^\ell_k} \right)^\star_{k \in \{1,\ldots,m\}}$ is the MLE of the vector $\left(\frac{\delta_{\mathcal{S}(k)}(P)}{c^\ell_k} \right)_{k \in \{1,\ldots,m\}}$ w.r.t. the matrix $A^{1,\ldots,\ell}$. The same holds for $\left(\frac{a^C_{\mathcal{S}(k)}(P)}{c^C_k} \right)^\star$ w.r.t. $A^{\ell+1,\ldots,d}$.

Proof. Since $P^\star \in \mathcal{E}_A$ we are able to use Proposition 12 to write every $P^\star(y_k)$ as

$$P^\star(y_k) = \frac{\left(a^B_{\mathcal{S}(k)}(P^\star)\right)\left(a^C_{\mathcal{S}(k)}(P^\star)\right)}{c^d_k\left(a^D_{\mathcal{S}(k)}(P^\star)\right)} = \frac{\left(\frac{a^B_{\mathcal{S}(k)}(P^\star)}{c^\ell_k}\right)\left(\frac{a^C_{\mathcal{S}(k)}(P^\star)}{c^C_k}\right)}{\frac{c^d_k}{c^\ell_k c^C_k}a^D_{\mathcal{S}(k)}(P^\star)}, \quad \forall y_k \in \mathcal{Y}.$$

In order for $\left(\frac{a^B_{\mathcal{S}(k)}(P^\star)}{c^\ell_k}\right)_{k\in\{1,\dots,m\}}$ to equal the MLE of $\left(\frac{a^B_{\mathcal{S}(k)}(P)}{c^\ell_k}\right)_{k\in\{1,\dots,m\}}$ w.r.t. $\mathcal{E}_{A^{1,\dots,\ell}}$ the following two conditions need to be satisfied

(1) $\left(\frac{a^B_{\mathcal{S}(k)}(P^\star)}{c^\ell_k}\right)_{k\in\{1,\dots,m\}} \in \mathcal{E}_{A^{1,\dots,\ell}}$

(2) $A^{1,\dots,\ell}\left(\frac{a^B_{\mathcal{S}(1)}(P^\star)}{c^\ell_1},\dots,\frac{a^B_{\mathcal{S}(m)}(P^\star)}{c^\ell_m}\right)^\intercal = A^{1,\dots,\ell}\left(\frac{a^B_{\mathcal{S}(1)}(P)}{c^\ell_1},\dots,\frac{a^B_{\mathcal{S}(m)}(P)}{c^\ell_m}\right)^\intercal.$

The last property above is also known as Birch's theorem and can be found for example in Corollary 7.3.9 in [Sullivant18]. For the proof of (1) we use the parameterization of $a^B_{\mathcal{S}(k)}(P^\star)$ in (6.5):

$$\frac{1}{c^\ell_k}a^B_{\mathcal{S}(k)}(P^\star) = \left(\prod_{i=1}^{\ell} t_{\mathcal{S}(i,k)}\right)\sum_{k'\in J^C(a^B_{\mathcal{S}(k)})}\frac{c^d_{k'}}{c^\ell_{k'}}\left(\prod_{i=\ell+1}^{d} t_{\mathcal{S}(i,k')}\right)$$

The sum is the same for every row in B that has a floret equal to $\mathcal{F}^C(a^B_{\mathcal{S}(k)})$. In other words, the sum is equal for every row in B that is connected to $a^D_{\mathcal{S}(k)} \in D$. Since D encodes the florets, the sum corresponds to a row in D. Therefore we are able to rewrite this sum depending on $a^D_{\mathcal{S}(k)}$ and considering that D lies in the rowspan of $A^{1,\dots,\ell}$, this results in $\frac{1}{c^\ell_k}a^B_{\mathcal{S}(k)}(P^\star) \in \mathcal{E}_{A^{1,\dots,\ell}}$.
(2) The second property directly follows from the definition of B and the fact that $P^\star$ is the MLE of P w.r.t $A^{1,\dots,d}$, hence

$$a^{A^{1,\dots,\ell}}_{\mathcal{S}(k)}(a^B_{\mathcal{S}(k)}(P^\star)) = a^{A^{1,\dots,\ell}}_{\mathcal{S}(k)}(P^\star) = a^{A^{1,\dots,\ell}}_{\mathcal{S}(k)}(P) = a^{A^{1,\dots,\ell}}_{\mathcal{S}(k)}(a^B_{\mathcal{S}(k)}(P))$$

The analogous result for the matrix C can be proven in the same manner. □

The next result provides us with a general formula for the MLE of a model that satisfies the GRIP.

Proposition 14. *Let A be a multipartition matrix with $d > 1$ partitions that satisfies the GRIP and let $F(a^\ell_{\mathcal{S}(\ell,k)})$ be the set of rows in A^ℓ belonging to the same floret with respect to $\biguplus_{i=1}^{\ell-1}A^i$. Then the MLE $P^\star$ of P has the following kth coordinate function*

$$P^\star(y_k) = \frac{a^1_{\mathcal{S}(1,k)}(P)}{c^d_k}\left(\prod_{i=2}^{d}\frac{a^i_{\mathcal{S}(i,k)}(P)}{\sum_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k)})}a^i_{j'}(P)}\right), \quad \forall y_k \in \mathcal{Y}.$$

Proof. Using Proposition 13 we are able to write

$$P^\star(y_k) = \frac{\left(\frac{a^B_{\mathcal{S}(k)}(P)}{c_k^{d-1}}\right)^\star \left(\frac{a^d_{\mathcal{S}(d,k)}(P)}{c_k^d}\right)^\star}{\frac{c_k^d}{c_k^{d-1}c_k^d} a^D_{\mathcal{S}(k)}(P)} = \frac{\left(\frac{a^B_{\mathcal{S}(k)}(P)}{c_k^{d-1}}\right)^\star \left(a^d_{\mathcal{S}(k)}(P)\right)}{\frac{c_k^d}{c_k^{d-1}} \sum\limits_{a^d_{j'} \in F(a^d_{\mathcal{S}(d,k)})} a^d_{j'}(P)}.$$

for all $y_k \in \mathcal{Y}$, $B = \Cup_{i=1}^{d-1} A^i$, $C = A^d$ and $D = B \cap C$. Here we use that the MLE in case of one partition is known and that $a^D_{\mathcal{S}(k)}$ can be written as the sum of elements in A^d that belong to the floret $a^D_{\mathcal{S}(k)}$ is associated to.

Let $B_\ell = \Cup_{i=1}^{\ell} A^i$. Since $\left(\frac{a^{B_{d-1}}_{\mathcal{S}(k)}(P)}{c_k^{d-1}}\right)^\star$ is the MLE with respect to $A^{1,\dots,d-1}$, we are able to apply Proposition 13 iteratively

$$\begin{aligned} P^\star(y_k) &= \frac{\left(\frac{a^{B_{d-1}}_{\mathcal{S}(k)}(P)}{c_k^{d-1}}\right)^\star \left(a^d_{\mathcal{S}(d,k)}(P)\right)}{\frac{c_k^d}{c_k^{d-1}} \sum\limits_{a^d_{j'} \in F(a^d_{\mathcal{S}(d,k)})} a_{j'}^{B_{d-1}}} \\ &= \frac{\left(\frac{a^{B_{d-2}}_{\mathcal{S}(k)}(P)}{c_k^{d-2}}\right)^\star \left(a^{d-1}_{\mathcal{S}(d-1,k)}(P)\right)\left(a^d_{\mathcal{S}(d,k)}(P)\right)}{\frac{c_k^d}{c_k^{d-2}} \sum\limits_{a^{d-1}_{j'} \in F(a^{d-1}_{\mathcal{S}(d-1,k)})} a^{d-1}_{j'} \sum\limits_{a^d_{j'} \in F(a^d_{\mathcal{S}(d,k)})} a^d_{j'}} \\ &= \frac{\left(\frac{a^{B_{d-3}}_{\mathcal{S}(k)}(P)}{c_k^{d-3}}\right)^\star \left(a^{d-2}_{\mathcal{S}(d-2,k)}(P)\right)\left(a^{d-1}_{\mathcal{S}(d-1,k)}(P)\right)\left(a^d_{\mathcal{S}(d,k)}(P)\right)}{\frac{c_k^d}{c_k^{d-3}} \sum\limits_{a^{d-2}_{j'} \in F(a^{d-2}_{\mathcal{S}(d-2,k)})} a^{d-2}_{j'} \sum\limits_{a^{d-1}_{j'} \in F(a^{d-1}_{\mathcal{S}(d-1,k)})} a^{d-1}_{j'} \sum\limits_{a^d_{j'} \in F(a^d_{\mathcal{S}(d,k)})} a^d_{j'}} \end{aligned}$$

Hence, this results in

$$P^\star(y_k) = \frac{a^1_{\mathcal{S}(1,k)}(P)}{c_k^d} \left(\prod_{i=2}^{d} \frac{a^i_{\mathcal{S}(i,k)}(P)}{\sum\limits_{a^i_{j'} \in F(a^i_{\mathcal{S}(i,k)})} a^i_{j'}(P)} \right), \quad \forall y_k \in \mathcal{Y}.$$

□

The previous proposition equips us with a formula for the MLE of a model with a representation that satsifies the GRIP. In order to prove Theorem 6.2.2 it remains to show that the iterative scaling algorithm results in exactly this formula.

Proof of Theorem 6.2.2. We prove this theorem per induction over the number of partition matrices d. Let $d = 1$, then the multipartition matrix only consists of one partition and the MLE is given by $P(y_k) = \frac{1}{c_k^d} a^1_{\mathcal{S}(i,k)(P)}$, as discussed in Section 2.5.2.

Let $P_\ell^\star(y_k)$ be the MLE of a distribution P w.r.t. $\mathcal{E}_{A^{1,\dots,\ell}}$ and let $A^{1,\dots,\ell+1}$ be a multi-partition matrix satisfying the GRIP, then $A^{1,\dots,\ell}$ also satisfies the GRIP. Hence, using the induction hypothesis and Proposition 13 allows us to write $P_\ell^\star$ in the following way

$$P_\ell^\star(y_k) = \frac{a^1_{\mathcal{S}(1,k)}(P)}{c_k^\ell}\left(\prod_{i=2}^{\ell}\frac{a^i_{\mathcal{S}(i,k)}(P)}{\sum\limits_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k)})} a^i_{j'}(P)}\right), \quad \forall y_k \in \mathcal{Y}.$$

Performing the next step of the iterative scaling algorithm, which consists of an e-projection of $P_\ell^\star$ to the linear family defined by $A^{\ell+1}$, results in

$$P_\ell^\star(y_k)\frac{a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P)}{a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P_\ell^\star)}, \tag{6.6}$$

for all $y_k \in \mathcal{Y}$. Therefore we take a closer look at $a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P_\ell^\star(y_k))$.

Since $P_{\ell+1}^\star$ is the MLE w.r.t. $\mathcal{E}_{A^{1,\dots,\ell+1}}$, we are able to use that $a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P) = a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P_{\ell+1}^\star)$. Writing $P_{\ell+1}^\star$ in the form proven in Proposition 13 leads to

$$\begin{aligned}
a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P) &= \sum_{k'\in I(a^{\ell+1}_{\mathcal{S}(\ell+1,k)})} \frac{a^1_{\mathcal{S}(1,k')}(P)}{c^{\ell+1}_{k'}}\left(\prod_{i=2}^{\ell+1}\frac{a^i_{\mathcal{S}(i,k')}(P)}{\sum\limits_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k')})} a^i_{j'}(P)}\right) \\
&= \frac{a^{\ell+1}_{\mathcal{S}(\ell+1,k)}}{\frac{c_k^{\ell+1}}{c_k^\ell}\sum\limits_{a^{\ell+1}_{j'}\in F(a^{\ell+1}_{\mathcal{S}(i,k)})} a^{\ell+1}_{j'}(P)} \sum_{k'\in I(a^{\ell+1}_{\mathcal{S}(\ell+1,k)})} \frac{a^1_{\mathcal{S}(1,k')}(P)}{c^\ell_{k'}}\left(\prod_{i=2}^{\ell}\frac{a^i_{\mathcal{S}(i,k')}(P)}{\sum\limits_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k')})} a^i_{j'}(P)}\right).
\end{aligned}$$

Hence we are able to conclude

$$\begin{aligned}
\frac{c_k^{\ell+1}}{c_k^\ell}\sum_{a^{\ell+1}_{j'}\in F(a^{\ell+1}_{\mathcal{S}(i,k)})} a^{\ell+1}_{j'}(P) &= \sum_{k'\in I(a^{\ell+1}_{\mathcal{S}(\ell+1,k)})} \frac{a^1_{\mathcal{S}(1,k')}(P)}{c^\ell_{k'}}\left(\prod_{i=2}^{\ell}\frac{a^i_{\mathcal{S}(i,k')}(P)}{\sum\limits_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k')})} a^i_{j'}(P)}\right) \\
&= a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P_\ell^\star(y_k)).
\end{aligned}$$

Now, we substitute $a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P_\ell^\star(y_k))$ in (6.6) in order to achieve the desired result

$$\begin{aligned}
P_\ell^\star(y_k)\frac{a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P)}{a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P_\ell^\star)} &= \frac{a^1_{\mathcal{S}(1,k)}(P)}{c_k^\ell}\left(\prod_{i=2}^{\ell}\frac{a^i_{\mathcal{S}(i,k)}(P)}{\sum\limits_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k)})} a^i_{j'}(P)}\right)\frac{a^{\ell+1}_{\mathcal{S}(\ell+1,k)}(P)}{\frac{c_k^{\ell+1}}{c_k^\ell}\sum\limits_{a^{\ell+1}_{j'}\in F(a^{\ell+1}_{\mathcal{S}(i,k)})} a^{\ell+1}_{j'}(P)} \\
&= \frac{a^1_{\mathcal{S}(1,k)}(P)}{c_k^{\ell+1}}\left(\prod_{i=2}^{\ell+1}\frac{a^i_{\mathcal{S}(i,k)}(P)}{\sum\limits_{a^i_{j'}\in F(a^i_{\mathcal{S}(i,k)})} a^i_{j'}(P)}\right).
\end{aligned}$$

□

6.3 Summary and Discussion of Chapter 6

This chapter discusses theoretical results related to the iterative scaling algorithm, also called iterative proportional scaling, and the *em*-algorithm.

At first, in Chapter 6.1, we apply the *em*-algorithm to find the MLE in a setting where latent variables are included in the model. There we explore three different methods to extend the latent space in a manner that takes the previously found lower dimensional local minimum into consideration. For the Integrated Information measure Φ_{CII} only the "safe" method, out of the three proposed ones, guarantees that the local minimum in the larger space is not worse compared to the one in the smaller space. The "natural" method makes use of a small constant c. In the examples in this chapter we compare different candidates and decide on $c = 0.01$, but the question remains how to choose this constant best. Maybe one should vary it through different runs of the algorithms. These approaches are a first step towards analyzing the connection among the different local minima that the *em*-algorithm converges to.

In the second section of this chapter, Section 6.2, we define a property that ensures one-cycle convergence of the iterative scaling algorithm. This property, GRIP, generalizes the known "Running Intersection Property" for hierarchical models. The definition of the corresponding models in case of the GRIP, called partition models, includes hierarchical models and staged tree models. Additionally, this new property has connections to algebraic statistics. These results are proven in [Coons24] and summarized in the Figure 6.11.

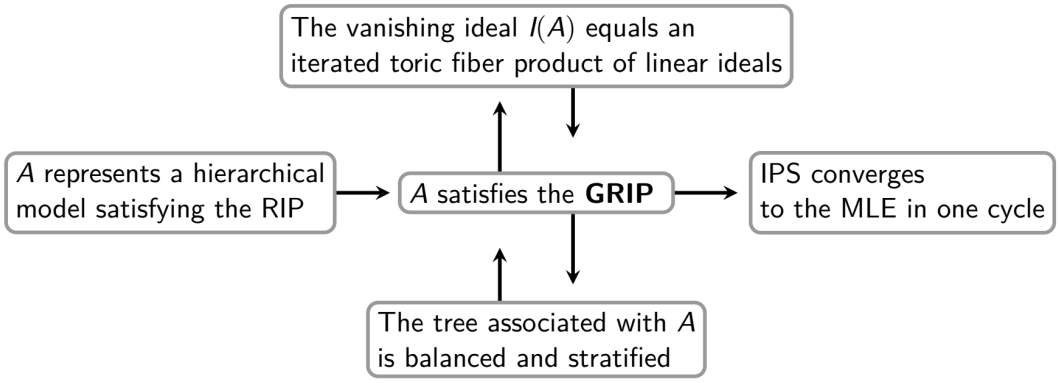


Figure 6.11: The main results from [Coons24] connecting the GRIP to hierarchical models, balanced and stratified staged trees and the iterated toric fiber product. This is a recreation of Figure 1 in [Coons24].

First, we can make use of the toric geometry of partition models. In [Coons24] we show that a multipartition matrix A satisfies the GRIP if and only if the vanishing ideal associated to A equals an iterated toric fiber product. The toric fiber product is an algebraic construction, introduced by Sullivant [Sullivant07], that enables us to build the overall model from smaller partition models. Hence, we use this method to prove the one-cycle convergence for models that satisfy the GRIP in [Coons24].

Additionally, the GRIP can be connected to models called "balanced and stratefied tree models". Staged tree models, also called chain event graphs in [Collazo18], are probability tree models that encode conditional independence relationships among events. They are connected to the GRIP in the following way: The tree associated with A is balanced and stratified if and only if A satisfies the GRIP.

Furthermore, we show that every hierarchical model that satisfies the RIP also satisfies the GRIP, hence it is a true generalization.

7 Conclusions and Outlook

Here we conclude this thesis and discuss possible directions for future research. This thesis is divided into four parts. First we discuss different information geometric measures related to the Integrated Information Theory. Next we apply one of these measures to simple simulated agents acting in a two-dimensional environment. Then we adapt the experimental setup such that the agents learn while being inside their environment. Finally, the last chapter holds more advanced theoretical results about the two information geometric algorithms that we apply in this thesis, namely the *em*- and the iterative scaling algorithm. Each of these chapters closes with a summary and discussion of its contents in its last section.

7.1 About Chapter 3

The applicability of the two measures that we propose in Chapter 3, namely the ground truth Integrated Information and the causal information integration, depends on whether a system has unknown exterior influences or not. In the case of known exterior influences the ground truth Integrated Information can be applied, which has a closed form solution. The calculation of the causal information integration for situations with unknown influences is more complicated and involves the *em*-algorithm. We define the split system in case of the causal information integration as the closure of the union of split systems for fixed sized latent spaces. This set still holds mathematical properties that remain to be analyzed, such as the question if it is finitely generated. Moreover, we conjecture that the causal information integration and the CIS Integrated Information measure are not equal solely based on numerical experiments. The mathematical proof of their relationship is an open problem.

7.2 About Chapter 4 and Chapter 5

The approaches in Chapter 4 and 5 are closely related as we analyze the information flow in acting and learning agents using a large number of information theoretic measures. Thereby we observe the impact of the different information flows on the behavior of the agents. In particular, we detect an antagonistic relationship between the concepts of Integrated Information and Morphological Computation. This result is in accordance with the view of proponents of the embodied artificial intelligence approach. In the introduction of [Hoffmann12] the authors stress the importance of the influence of embodiment:

> "While embodiment has often been used in its trivial meaning, i.e. 'intelligence requires a body', there are deeper and more important consequences, concerned with connecting brain, body, and environment. The behavior of any system is not merely the outcome of an internal control structure (such as the central nervous system); it is also affected by the ecological niche in which the system is physically embedded, by its morphology (the shape of its body and limbs, as well as the type and placement of sensors and effectors), and by the material properties of the elements composing the morphology."

In this thesis we are additionally able to observe that the capabilities of the controller influence the interaction of the body with the environment. A control structure with an accurate world model can facilitate a high Morphological Computation that might not be achievable for agents with a less complex structure, provided that the task has to be learned and is not inherited. We now take a closer look at the relationship between Morphological Computation and Integrated Information.

7.2.1 Relationship between Morphological Computation and Integrated Information

The Chapters 5 and 4 are concerned with the information flow in simulated agents, faced with similar tasks. In both cases we are especially interested in the relationship between the concepts of Morphological Computation and the controller complexity, including Integrated Information. Therefore, we now summarize the main results from both chapters and relate them to each other.

Here we especially consider five key aspects of the a situation an agent can be in, namely its world model, action effect, Integrated Information, effective information integration and Morphological Computation. We distinguish between three situations, specifically an inaccurate world model, an updating world model and an accurate world model. Therefore, we differentiate between the two components of the controller complexity, which are the complexity of the internal world model and the Integrated Information, here in this way.

The agents in Chapter 4 do not alter or even indirectly influence their world model. There they have access to the empirical world model, which was sampled for every sensor length beforehand. This directly influences the Morphological Computation, which depends on the empirical world model. In that context we first observe the antagonistic relationship between Integrated Information and Morphological Computation, given in Figure 4.27. In cases with high Morphological Computation the Integrated Information is low and vice versa.

However, the effect the agent has on their next sensory state is very low overall, as one can see in Figure 4.5, such that the controller has a limited influence on the success of the agent in general. For short sensors, this is further aggravated by the architecture of these agents. Here they have to take two steps before information processed in the controller can reach the actuators. For short sensors it is necessary to directly react to sensory data and therefore the high Integrated Information in the case of shorter sensors has no influence on the behavior of the agent. This leads to the effective information integration being close to zero. In the case of longer sensors the effective information integration is still very low, as depicted in Figure 4.30, and therefore we are not able to draw further conclusions about the behavior of this measure in relation to the Morphological Computation here.

The experiments in Chapter 5 enrich the picture further. There we analyze the information flow of agents that have to learn while moving through their environment and the information processing in the controller is one step faster. First, we discuss the results for the ideal agents that have access to their perfectly accurate world models. They can indirectly affect these models through their actions. We observe in Figure 5.13 that successful agents have a much higher Morphological Computation compared to the unsuccessful ones. The action effect, on the other hand, is significantly larger for the unsuccessful ideal agents. Hence, in the case of high Morphological Computation the agents better interact with their environment, which leads to a more stable situation in which their actions do not influence the outcome of their next movement as much. Even though the unsuccessful agents move in a way that leads to less Morphological Computation their world model is still accurate and there is next to no effective information integration.

When we vary the accuracy of the world model, described in Section 5.6.2, we observe that the effective information integration is higher for worse world models, as we can see in Figure 5.15. Similarly, the actions of an agent have a higher impact for the more inaccurate world models, depicted in Figure 5.14. In these situations the Morphological Computation has a lower value.

Now we highlight a possible intuition behind these mechanisms by considering the example of a child on a bike from the introduction in Section 5.1. When a child starts to

learn how to ride a bike its expectations are inaccurate, hence the world model does not fit to the situation. It tries to ride the bike slowly in the beginning and can not utilize the dynamics of the environment. Although the bike is not part of the body of the agent we use this analogy to symbolize a low Morphological Computation in our setting. Each small movement of the child leads to a strong reaction of the bike and potential failure of the task, the effect of its actions are high. At the same time the child needs to concentrate and pay attention to every movement, which we relate to a high Integrated Information. This situation is symbolized on the left in Figure 7.1.

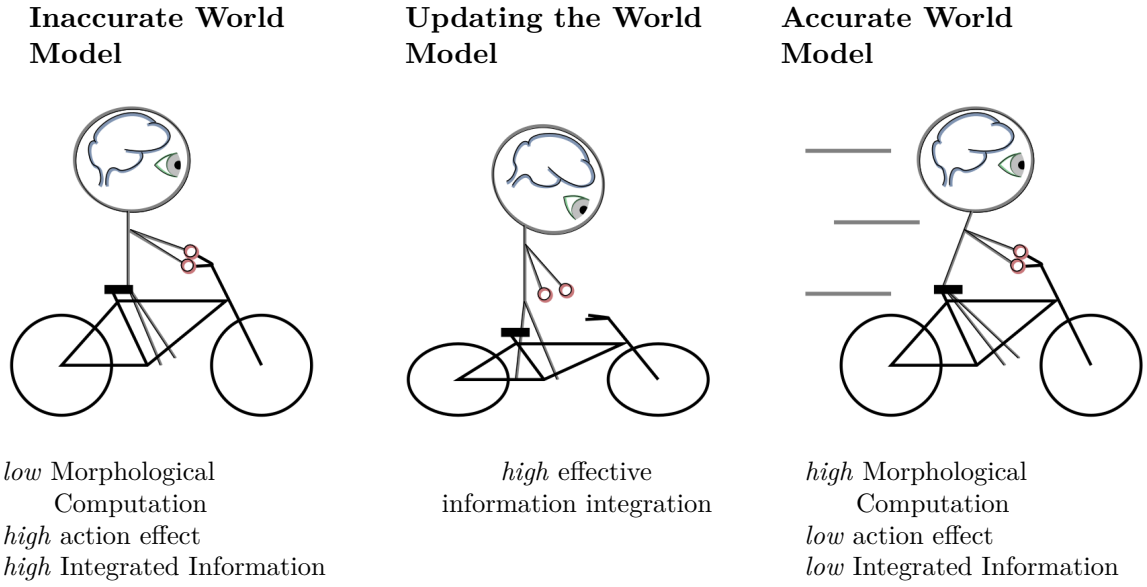


Figure 7.1: Visualization of the different situations of an agent depending on its world model.

While failing the child constantly learns and updates its own world model. As long as the world model does not accurately describe the dynamics of the world a high effective information integration is required, as discussed in the context of Figure 5.15. The middle column of Figure 7.1 depicts the learning process. This stage does not exist in the case of the ideal agents, since they have access to their empirical world models and do not need to learn them. Once the world model is accurate, the connection between speed and stability is understood and the child succeeds, as depicted on the right of Figure 7.1. The child now utilizes the dynamics of the world and the process becomes much more stable, which results in a high Morphological Computation. At a higher speed, the movements do not need to be as exact and the child does not have to be as attentive, compared to before. This translates to a low action effect and a low Integrated Information value.

The same relationships as described above can be observed in case of the fully connected agents, which have to additionally learn their internal world models. In this case we are not able to measure the accuracy of the world model directly, but have to rely on the connection of the world model to the Morphological Computation. Here the successful agents have a higher Morphological Computation and lower action effect, compared to the unsuccessful ones, as depicted in Figure 5.18. Additionally, the Integrated Information and effective information integration is lower for successful, compared to unsuccessful agents. Therefore, the successful agents relate to the stage in Figure 7.1 with the accurate world model, whereas the results of the measures for unsuccessful agents suggest an inaccurate world model as depicted to the left of Figure 7.1.

The high value of Integrated Information for unsuccessful agents is not just a byproduct of the relationship between Morphological Computation and Integrated Information, but necessary in order to update the world model. We are able to substantiate this claim by considering split agents, which are not able to integrate information in their controller. These are almost not able to perform the task with a reasonable success rate and the few that do integrate different information sources directly in their world model, as discussed in the context of Figure 5.21. Hence, the learning stage in the middle of Figure 7.1 is not possible without integrating information, even in such a simple task like the one presented here.

The results from Chapter 4 only partly relate to the Figure above, because in that case the agents all have an accurate world model. The only variation lies in the morphology of the agents. For longer sensors the situation of the agents relate to the column on the right of Figure 7.1. The agents perform well with a high Morphological Computation and low Integrated Information. If the sensors are small, then this belongs to the situation described in the left of Figure 7.1, except for the action effect, which is always low because of the nature of this experimental setup. Furthermore, since the agents are not able to influence their own world models, no part of the experiments from Chapter 4 relate to the middle of Figure 7.1.

The figure above illustrates the different stages of an agent by classifying them depending on the state of the world model. Note that these stages are not as strict as they appear in that figure, but the transitions between an inaccurate, updating and accurate world model are fluid. Furthermore, this example should only give an intuition to the mechanisms described by the various information theoretic measures. We do not claim that a child experiences exactly these processes, nor that an agent that updates its world model starts to understand the task.

7.2.2 Potential Future Research Directions

This approach allows us to analyze the situation thoroughly, but it has the caveat that it is only applicable for discrete and very small networks at the moment. For large networks the involved conditional probability distributions become intractable.

Analyzing more than one agent at time could lead to interesting research questions and observations without increasing the size of the discussed networks. There one could analyze to which degree interacting agents cooperate, more precisely, to which degree the information flows in their controllers depend on each other. Furthermore, are we able to detect solely from the information flows whether these agents cooperate or oppose each other?

Another goal for future research is to extend this framework to more involved simulated agents, tasks or environments. Similarly, the learning algorithms discussed in this thesis were chosen because of their intuitive geometric interpretation and their close relationship to the rest of the framework. It would be interesting to test how more efficient learning algorithms influence the information flows inside the agents. This could lead to new insides to the behavior of embodied agents. Although, when dealing with more complicated situations the mechanisms observed in this thesis might get overshadowed by stronger influences.

In particular, the translation of these measures to the continuous setting is a challenge, which would lead to wide range of potential areas of application. In that case information theoretic quantities, such as entropy or the KL-divergence, can rarely be calculated directly but need to be estimated. Despite these additional challenges it might be beneficial to extend the framework to the continuous setting, since the human brain is not a discrete system. Therefore it might be necessary to analyze continuous variables in order to advance the understanding of human cognition.

An ongoing cooperation with Prof. Dr. Verena V. Hafner and Yasmin K. Georgie aims at formulating the ideas of the information theoretic measures, discussed in this thesis, for a variational auto-encoder. They are interested in building a minimal self in a robotic setting, as described in [Nguyen21; Georgie19]. In this collaboration the variational auto-encoder is the control architecture for a real humanoid robot that performs a task related to the sense of agency. In this setting information theoretic measures could be used to analyze some properties of the latent space of the system. Additionally, we would be able to observe the impact of real embodiment on these measures and we would additionally be able to relate the integration of the different sensor and actuator modalities inside the system to its success. We would imagine that the control behavior of an agent that understands the interplay between its visual signals and joint positions, for example, is more robust compared to an agent that considers all the modalities as independent of each other.

7.3 About Chapter 6

The theoretical results regarding the *em*-algorithm in Section 6.1 are only a first step towards analyzing the behavior of the local minima the algorithm converges to. In the situation with latent variables, discussed in Section 2.5.1, Amari et al. show in [Amari92] that the global minimum in the visible space is equal to the global minimum in the larger space. However, the question remains if the same holds in case the algorithm converges to a local minimum.

In Section 6.2 we discuss partition models, which are a generalization of hierarchical models. We show that a partition model for which the associated matrix satisfies the GRIP forms the MLE in one cycle in a particular way. Each step results in the MLE of the model associated to the partitions the algorithm already considers. Now, one could ask whether the reverse statement is also true. Does every model with an MLE with this structure have a matrix representation satisfying the GRIP? Additionally, there is a stronger version of the running intersection property for hierarchical models guaranteeing one-cycle convergence independently of the order of the partitions. One remaining question is whether there exists such a strong version of the GRIP for partition models.

In conclusion, this thesis discusses an information theoretic approach for the thorough analysis of the different information flows in an artificial agent. This framework was able to confirm previously made intuitive hypotheses in a simplistic example and remains to be translated to more involved settings.

Bibliography

[Aaronson14a] S. Aaronson. *Giulio Tononi and Me: A Phi-nal Exchange*. 2014. URL: https://scottaaronson.blog/?p=1823 (visited on 07/10/2023).

[Aaronson14b] S. Aaronson. *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*. 2014. URL: https://www.scottaaronson.com/blog/?p=1799 (visited on 07/10/2023).

[Albantakis14] L. Albantakis, A. Hintze, C. Koch, C. Adami, and G. Tononi. "Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity". In: *PLOS Computational Biology* 10.12 (2014), pp. 1–19. DOI: 10.1371/journal.pcbi.1003966.

[Albantakis15] L. Albantakis and G. Tononi. "The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats". In: *Entropy* (2015), pp. 5472–5502. DOI: 10.3390/e17085472.

[Albantakis18] L. Albantakis. "A Tale of Two Animats: What Does It Take to Have Goals?" In: *Wandering Towards a Goal: How Can Mindless Mathematical Laws Give Rise to Aims and Intention?* Ed. by A. Aguirre, B. Foster, and Z. Merali. Cham: Springer International Publishing, 2018, pp. 5–15. ISBN: 978-3-319-75726-1. DOI: 10.1007/978-3-319-75726-1_2.

[Albantakis19] L. Albantakis and G. Tononi. "Causal Composition: Structural Differences among Dynamically Equivalent Systems". In: *Entropy* 21.10 (2019). DOI: 10.3390/e21100989.

[Albantakis22] L. Albantakis, L. Barbosa, G. Findlay, M. Grasso, A. M. Haun, W. Marshall, W. G. Mayner, A. Zaeemzadeh, M. Boly, B. E. Juel, S. Sasai, K. Fujii, I. David, J. Hendren, J. P. Lang, and G. Tononi. *Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms*. 2022. DOI: 10.48550/ARXIV.2212.14787.

[Albantakis23] L. Albantakis, R. Prentner, and I. Durham. *Measuring the integrated information of a quantum mechanism*. 2023. DOI: 10.48550/ARXIV.2301.02244.

[Amari16] S. Amari. *Information Geometry and Its Applications*. Tokyo: Springer, 2016. ISBN: 978-4-431-55977-1.

[Amari18] S. Amari, N. Tsuchiya, and M. Oizumi. "Geometry of Information Integration". In: *Information Geometry and Its Applications*. Ed. by N. Ay, P. Gibilisco, and F. Matúš. Cham: Springer, 2018, pp. 3–17. ISBN: 978-3-030-07405-0.

[Amari85] S. Amari. *Differential-Geometrical Methods in Statistics*. New York: Springer, 1985. ISBN: 978-0-387-96056-2.

[Amari92] S. Amari, K. Kurata, and H. Nagaoka. "Information geometry of Boltzmann machines". In: *IEEE Transactions on Neural Networks* (1992), pp. 260–271. DOI: 10.1109/72.125867.

[Amari95] S. Amari. "Information Geometry of the EM and em Algorithms for Neural Networks". In: *Neural Networks* (1995), pp. 1379–1408. DOI: 10.1016/0893-6080(95)00003-8.

[Andersson97] S. A. Andersson, D. Madigan, and M. D. Perlman. "On the Markov Equivalence of Chain Graphs, Undirected Graphs, and Acyclic Digraphs". In: *Scandinavian Journal of Statistics* 24.1 (1997), pp. 81–102. DOI: 10.1111/1467-9469.00050.

[Athreya06] K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer Texts in Statistics. New York: Springer, 2006. ISBN: 978-0-387-96056-2.

[Attias03] H. Attias. "Planning by Probabilistic Inference". In: *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. 2003, pp. 9–16.

[Autumn02] K. Autumn, M. Sitti, Y. A. Liang, A. M. Peattie, W. R. Hansen, S. Sponberg, T. W. Kenny, R. Fearing, J. N. Israelachvili, and R. J. Full. "Evidence for van der Waals adhesion in gecko setae". In: *Proceedings of the National Academy of Sciences* 99.19 (2002), pp. 12252–12256. DOI: 10.1073/pnas.192252799.

[Ay01] N. Ay. "Information Geometry on Complexity and Stochastic Interaction". In: *MPI MIS PREPRINT 95* (2001). URL: https://www.mis.mpg.de/preprints/2001/preprint2001_95.pdf (visited on 04/28/2022).

[Ay08a] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbricht. "Predictive information and explorative behavior of autonomous robots". In: *European Physical Journal B* (2008), pp. 329–339. DOI: 10.1140/epjb/e2008-00175-0.

[Ay08b] N. Ay and D. Polani. "Information flows in causal networks". In: *Advances in complex systems* 11.1 (2008), pp. 17–41. ISSN: 0219-5259. DOI: 10.1142/S0219525908001465.

[Ay13a] N. Ay and K. Zahedi. "An Information Theoretic Approach to Intention and Deliberative Decision-Making of Embodied Systems". In: *Advances in cognitive neurodynamics III*. Heidelberg: Springer, 2013.

[Ay13b] N. Ay and K. Zahedi. "Causal Effects for Prediction and Deliberative Decision Making of Embodied Systems". English. In: *Advances in Cognitive Neurodynamics (III)*. Ed. by Y. Yamaguchi. Dordrecht: Springer, 2013, pp. 499–506. ISBN: 978-94-007-4791-3.

[Ay14] N. Ay and K. Zahedi. "On the Causal Structure of the Sensorimotor Loop". In: *Guided Self-Organization: Inception*. Ed. by M. Prokopenko. Berlin, Heidelberg: Springer, 2014, pp. 261–294. ISBN: 978-3-642-53733-2. DOI: 10.1007/978-3-642-53734-9_9.

[Ay15a] N. Ay. "Information Geometry on Complexity and Stochastic Interaction". In: *Entropy* 17 (2015), pp. 2432–2458. DOI: 10.3390/e17042432.

[Ay15b] N. Ay and W. Löhr. "The Umwelt of an embodied agent-a measure-theoretic definition". In: *Theory in Biosciences* 134 (Dec. 2015), pp. 105–116. DOI: 10.1007/s12064-015-0217-3.

[Ay17] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information Geometry*. Cham: Springer, 2017. ISBN: 978-3-319-56477-7.

[Ay20] N. Ay, D. Polani, and N. Virgo. "Information decomposition based on cooperative game theory". In: *Kybernetika* (2020), pp. 979–1014. ISSN: 0023-5954, 1805-949X. DOI: 10.14736/kyb-2020-5-0979.

[Ay21] N. Ay. "Confounding Ghost Channels and Causality: A New Approach to Causal Information Flows". In: *Vietnam Journal of Mathematics* (2021), pp. 547–576. DOI: 10.1007/s10013-021-00511-w.

[Aziz20] S. Aziz, A. S. Rambely, K. B. Gan, and W. R. Wan Din. "Kinetics Study in Parachute Landing Fall Technique by Comparing Professional and Amateur Malaysian Army Parachutists Using Kane's Method". In: *Mathematics* 8.6 (2020), pp. 1–15. ISSN: 2227-7390. DOI: 10.3390/math8060917.

[Baldini22] P. Baldini. *Reservoir Computing in robotics: a review*. 2022. DOI: 10.48550/ARXIV.2206.11222.

[Balduzzi08] D. Balduzzi and G. Tononi. "Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework". In: *PLoS Computational Biology* 6 (2008). DOI: 10.1371/journal.pcbi.1000091.

[Barbosa20] L. S. Barbosa, W. Marshall, S. Streipert, L. Albantakis, and G. Tononi. "A measure for intrinsic information". In: *Scientific Reports* 10.18803 (2020). DOI: 10.1038/s41598-020-75943-4.

[Barbosa21] L. S. Barbosa, W. Marshall, L. Albantakis, and G. Tononi. "Mechanism Integrated Information". In: *Entropy* 23 (2021). ISSN: 1099-4300. DOI: 10.3390/e23030362.

[Barrett11] A. B. Barrett and A. K. Seth. "Practical Measures of Integrated Information for Time-Series Data". In: *PLOS Computational Biology* 7.1 (2011), pp. 1–18. DOI: 10.1371/journal.pcbi.1001052.

[Bayne18] T. Bayne. "On the axiomatic foundations of the integrated information theory of consciousness". In: *Neuroscience of Consciousness* (2018). DOI: 10.1093/nc/niy007.

[Ben-Gal08] I. Ben-Gal. "Bayesian Networks". In: *Encyclopedia of Statistics in Quality and Reliability*. John Wiley & Sons, Ltd, 2008. ISBN: 978-0-470-06157-2. DOI: https://doi.org/10.1002/9780470061572.eqr089.

[Berman20] R. J. Berman. "The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations". In: *Numerische Mathematik* 145 (2020), pp. 771–836. DOI: 10.1007/s00211-020-01127-x.

[Bernstein67] N. Bernstein. *The Co-ordination and Regulation of Movements*. Oxford: Pergamon Press, 1967. ISBN: 978-0-080-11940-3.

[Bialek01] W. Bialek, I. Nemenman, and N. Tishby. "Predictability, Complexity, and Learning". In: *Neural Comput.* 13.11 (Nov. 2001), pp. 2409–2463. ISSN: 0899-7667. DOI: 10.1162/089976601753195969.

[Borgelt09] C. Borgelt, M. Steinbrecher, and R. R. Kruse. *Graphical models: Representations for learning, reasoning and data mining*. Wiley Series in Computational Statistics. Wiley, 2009. ISBN: 978-0-470-72210-7.

[Bourgine04] P. E. Bourgine and J. Stewart. "Autopoiesis and Cognition". In: *Artificial Life* 10 (2004), pp. 327–345. DOI: 10.1162/1064546041255557.

[Braitenberg84] V. Braitenberg. *Vehicles: Experiments in synthetic psychology*. Cambridge: MIT Press, 1984. ISBN: 978-0-262-52112-3.

[Braun09] D. Braun, A. Aertsen, D. Wolpert, and C. Mehring. "Learning Optimal Adaptation Strategies in Unpredictable Motor Tasks". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29 (June 2009), pp. 6472–6478. DOI: 10.1523/JNEUROSCI.3075-08.2009.

[Brémaud17] P. Brémaud. *Discrete Probability Models and Methods*. Cham: Springer, 2017. ISBN: 978-3-319-43475-9.

[Brooks86] R. Brooks. "A robust layered control system for a mobile robot". In: *IEEE Journal on Robotics and Automation* 2.1 (1986), pp. 14–23. DOI: 10.1109/JRA.1986.1087032.

[Brooks91a] R. A. Brooks. "Intelligence without reason". In: *Proceedings of the 12th international joint conference on artificial intelligence (IJCAI-91)*. Ed. by R. R. Myopoulos J. Morgan Kaufmann publishers Inc., 1991, pp. 569–595.

[Brooks91b] R. A. Brooks. "Intelligence without representation". In: *Artificial Intelligence* 47.1 (1991), pp. 139–159. ISSN: 0004-3702. DOI: 10.1016/0004-3702(91)90053-M.

[Brown93] J. B. Brown, P. J. Chase, and A. O. Pittenger. "Order Independence and Factor Convergence in iterative Scaling". In: *Linear Algebra and its Applications* 190 (1993), pp. 1–38. DOI: doi.org/10.1016/0024-3795(93)90218-D.

[Casali13] A. G. Casali, O. Gosseries, M. Rosanova, M. Boly, S. Sarasso, K. R. Casali, S. Casarotto, M.-A. Bruno, S. Laureys, G. Tononi, and M. Massimini1. "A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior". In: *Science translational medicine* 198ra105 (2013). DOI: 10.1126/scitranslmed.3006294.

[Cerullo15] M. A. Cerullo. "The Problem with Phi: A Critique of Integrated Information Theory". In: *PLOS Computational Biology* 9 (2015). DOI: 10.1371/journal.pcbi.1004286.

[Choromanska15] A. Choromanska, M. Henaff, M. Mathieu, G. Arous, and Y. LeCun. "The Loss Surfaces of Multilayer Networks". In: *Proceedings of Machine Learning Research* 38 (2015), pp. 192–204.

[Collazo18] R. A. Collazo, C. Görgen, and J. Smith. *Chain Event Graphs*. Boca Raton, FL: CRC Press, 2018. ISBN: 978-0-367-57231-0.

[Conant70] R. C. Conant and W. R. Ashby. "Every good regulator of a system must be a model of that system". In: *International Journal of Systems Science* 1.2 (1970), pp. 89–97. DOI: 10.1080/00207727008920220.

[Coons21] J. I. Coons and S. Sullivant. "Quasi-independence models with rational maximum likelihood estimator". In: *Journal of Symbolic Computation* 104 (2021), pp. 917–941. DOI: 10.1016/j.jsc.2020.10.006.

[Coons24] J. I. Coons, C. Langer, and M. Ruddy. "Classical iterative proportional scaling of log-linear models with rational maximum likelihood estimator". In: *International Journal of Approximate Reasoning* 164 (2024). DOI: 10.1016/j.ijar.2023.109043.

[Cooper10] R. Cooper. "Forward and Inverse Models in Motor Control and Cognitive Control". In: *Proceedings of the International Symposium on AI Inspired Biology - A Symposium at the AISB 2010 Convention*. 2010, pp. 108–110. ISBN: 1-902-95692-3.

[Cover06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 2006. ISBN: 978-0-471-06259-2.

[Csiszár04] I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial*. Vol. 1. 4. 2004, pp. 417–528. DOI: 10.1561/0100000004.

[Csiszár75] I. Csiszár. "I-divergence geometry of probability distributions and minimization problems". In: *The Annals of Probability* 3.1 (1975), pp. 145–158. DOI: 10.1214/aop/1176996454.

[Csiszár84] I. Csiszár and G. Tsunády. "Information geometry and alternating minimization procedures". In: *Statistics and Decisions (Supplementary Issue, No.1)*. Ed. by e. a. Dedewicz E.F. Munich: Oldenburg Verlag, 1984, pp. 205–237.

[Csiszár89] I. Csiszár. "A geometric interpretation of Darroch and Ratcliff's Generalized Iterative Scaling". In: *The Annals of Statistics* 17.3 (1989), pp. 1409–1413. DOI: 10.1214/aos/1176347279.

[Darroch72] J. Darroch and D. Ratcliff. "Generalized iterative scaling for log-linear models". In: *The Annals of Mathematical Statistics* 43 (1972), pp. 1470–1480. DOI: 10.1214/aoms/1177692379.

[Darwiche09] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. 1st. Cambridge: Cambridge University Press, 2009. ISBN: 0-521-88438-1.

[Dawid79] A. Dawid. "Conditional Independence in Statistical Theory". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41 (1979), pp. 1–15. DOI: 10.1111/j.2517-6161.1979.tb01052.x.

[Deming40] W. Deming and F. Stephan. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known". In: *The Annals of Mathematical Statistics* 11 (1940), pp. 427–444. DOI: 10.1214/aoms/1177731829.

[Dempster77] A. Dempster, N. Laird, and D. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society* 39 (1977), pp. 2–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[Der12] R. Der and G. Martius. *The Playful Machine - Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Vol. 15. Berlin, Heidelberg: Springer, 2012. ISBN: 978-3-642-20252-0. DOI: 10.1007/978-3-642-20253-7.

[Diamond88] J. Diamond. "Why cats have nine lives". In: *Nature* 332 (1988), pp. 586–587. DOI: 10.1038/332586a0.

[Drton09a] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*. Basel: Birkhäuser, 2009. ISBN: 978-3-7643-8904-8.

[Drton09b] M. Drton. "Discrete chain graph models". In: *Bernoulli* 15.3 (2009), pp. 736–753. ISSN: 13507265. URL: http://www.jstor.org/stable/20680175.

[Duarte21] E. Duarte, O. Marigliano, and B. Sturmfels. "Discrete statistical models with rational maximum likelihood estimator". In: *Bernoulli* 27.1 (2021), pp. 135–154. DOI: 10.3150/20-bej1231.

[Edlund11] J. A. Edlund, N. Chaumont, A. Hintze, C. Koch, G. Tononi, and C. Adami. "Integrated Information Increases with Fitness in the Evolution of Animats". In: *PLoS Computational Biology* 7.10 (2011). DOI: 10.1371/journal.pcbi.1002236.

[Franceschini92] N. Franceschini, J.-M. Pichon, and C. Blanes. "From Insect Vision to Robot Vision". In: *Philosophical Transactions of The Royal Society B Biological Sciences* 337 (1992), pp. 283–294. DOI: 10.1098/rstb.1992.0106.

[Francis76] B. Francis and W. Wonham. "The Internal Model Principle of Control Theory". In: *Automatica* 12 (1976), pp. 457–465. DOI: 10.1016/0005-1098(76)90006-6.

[Frydenberg90] M. Frydenberg. "The Chain Graph Markov Property". In: *Scandinavian Journal of Statistics* (1990), pp. 333–353.

[Füchslin13] R. Füchslin, A. Dzyakanchuk, D. Flumini, H. Hauser, K. Hunt, R. Luchsinger, B. Reller, S. Scheidegger, and R. Walker. "Morphological computation and morphological control: steps toward a formal theory and applications". In: *Artificial Life* (2013), pp. 9–34. DOI: 10.1162/ARTL_a_00079.

[Gal00] S. Gallagher. "Philosophical conceptions of the self: implications for cognitive science". In: *Trends in Cognitive Sciences* 4.1 (2000), pp. 14–21. ISSN: 1364-6613. DOI: 10.1016/S1364-6613(99)01417-5.

[Geiger01] D. Geiger, D. Heckerman, H. King, and C. Meek. "Stratified Exponential Families: Graphical Models and Model Selection". In: *The Annals of Statistics* (2001), pp. 505–529. DOI: 10.1214/aos/1009210550.

[Geiger98] D. Geiger and C. Meek. "Graphical models and exponential families". In: *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence* (1998), pp. 156–165. DOI: 10.5555/2074094.2074113.

[Gell-Mann99] M. Gell-Mann. "Simplicity and Complexity in the Description of Nature". In: *Engineering and Science* 51 (1999), pp. 2–9. ISSN: 0013-7812.

[Georgie19] Y. K. Georgie, G. Schillaci, and V. V. Hafner. "An interdisciplinary overview of developmental indices and behavioral measures of the minimal self". In: *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2019, pp. 129–136. DOI: 10.1109/DEVLRN.2019.8850703.

[Ghazi-Zahedi10] K. Zahedi, N. Ay, and R. Der. "Higher Coordination With Less Control—A Result of Information Maximization in the Sensorimotor Loop". In: *Adaptive Behavior* 18.3-4 (2010), pp. 338–355. DOI: 10.1177/1059712310375314.

[Ghazi-Zahedi13] K. Ghazi-Zahedi and N. Ay. "Quantifying Morphological Computation". In: *Entropy* 15.5 (2013), pp. 1887–1915. DOI: 10.3390/e15051887.

[Ghazi-Zahedi16] K. Ghazi-Zahedi, D. Häufle, G. Montúfar, S. Schmitt, and N. Ay. "Evaluating Morphological Computation in Muscle and DC-Motor Driven Models of Hopping Movements". In: *Frontiers in Robotics and AI* 2 (2016). DOI: 10.3389/frobt.2016.00042.

[Ghazi-Zahedi17a] K. Ghazi-Zahedi, C. Langer, and N. Ay. "Morphological Computation: Synergy of Body and Brain". In: *Entropy* 19.456 (2017). DOI: 10.3390/e19090456.

[Ghazi-Zahedi17b] K. Ghazi-Zahedi, R. Deimel, G. Montúfar, V. Wall, and O. Brock. "Morphological Computation: The Good, the Bad, and the Ugly". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2017, pp. 464–469. DOI: 10.1109/IROS.2017.8202194.

[Ghazi-Zahedi19] K. Ghazi-Zahedi. *Morphological Intelligence*. Cham: Springer, 2019. ISBN: 978-3-030-20620-8.

[Goodman02] J. Goodman. "Sequential Conditional Generalized Iterative Scaling". In: *In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*. 2002, pp. 9–16. DOI: 10.3115/1073083.1073086.

[Grassberger86] P. Grassberger. "Toward a quantitative theory of self-generated complexity". In: *International Journal of Theoretical Physics* 25 (Jan. 1986), pp. 907–938. DOI: 10.1007/BF00668821.

[Haberman74] S. J. Haberman. *The Analysis of Frequency Data*. Chicago: University of Chicago Press, 1974. ISBN: 0-226-31184-8.

[Hafner20] V. V. Hafner, P. Loviken, A. Pico Villalpando, and G. Schillaci. "Prerequisites for an Artificial Self". In: *Frontiers in Neurorobotics* 14 (2020). ISSN: 1662-5218. URL: https://www.frontiersin.org/articles/10.3389/fnbot.2020.00005 (visited on 05/22/2023).

[Harder] M. Harder, C. Salge, and D. Polani. "A Bivariate Measure of Redundant Information". In: *Physical Review E* 87.1 (). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.87.012130.

[Häufle10] D. Häufle, S. Grimmer, and A. Seyfarth. "The role of intrinsic muscle properties for stable hopping—stability is achieved by the force–velocity relation". In: *Bioinspiration & Biomimetics* 5.1 (Feb. 2010). DOI: 10.1088/1748-3182/5/1/016004.

[Häufle14] D. Häufle, M. Günther, G. Wunner, and S. Schmitt. "Quantifying control effort of biological and technical movements: An information-entropy-based approach". In: *Physical Review E* 89 (Jan. 2014). DOI: 10.1103/PhysRevE.89.012716.

[Haun19] A. Haun and G. Tononi. "Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience". In: *Entropy* 21.1160 (2019). DOI: 10.3390/e21121160.

[Heider44] F. Heider and M. Simmel. "An Experimental Study of Apparent Behavior". In: *The American Journal of Psychology* 57.2 (1944), pp. 243–259. ISSN: 00029556. URL: http://www.jstor.org/stable/1416950 (visited on 02/09/2023).

[Hertz91] J. Hertz, A. Krogh, and R. Palmer. *Introduction to the Theory of Neural Computation*. Redwood City: Addison-Wesley Publishing Company, 1991. ISBN: 0-201-50395-6.

[Hinton83] G. Hinton and T. Sejnowski. "Optimal perceptual inference". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1983, pp. 448–453.

[Hoel16] E. Hoel, L. Albantakis, W. Marshall, and G. Tononi. "Can the macro beat the micro? Integrated information across spatiotemporal scales". In: *Neuroscience of Consciousness* 2016 (2016). DOI: 10.1093/nc/niw012.

[Hoffmann12] M. Hoffmann and R. Pfeifer. *The implications of embodiment for behavior and cognition: animal and robotic case studies*. 2012. DOI: 10.48550/arXiv.1202.0440.

[Hoffmann17] M. Hoffmann and V. Müller. "Simple or Complex Bodies? Trade-offs in Exploiting Body Morphology for Control". In: *Studies in Applied Philosophy, Epistemology and Rational Ethics* (2017), pp. 335–345. DOI: 10.1007/978-3-319-43784-2_17.

[Hopfield82] J. J. Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 79. 8. Apr. 1982, pp. 2554–2558. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC346238/ (visited on 07/13/2023).

[Ising25] E. Ising. "Beitrag zur Theorie des Ferromagnetismus". In: *Zeitschrift für Physik* 31 (1925), pp. 253–258. DOI: 10.1007/BF02980577.

[Jaynes57] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *The Physical Review* 106 (1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[Jordan96] M. I. Jordan. "Chapter 2 Computational aspects of motor control and motor learning". In: *Motor skills*. Ed. by H. Heuer and S. W. Keele. Vol. 2. Handbook of Perception and Action. Academic Press, 1996, pp. 71–120. DOI: https://doi.org/10.1016/S1874-5822(06)80005-8.

[Kanwal17] M. S. Kanwal, J. A. Grochow, and N. Ay. "Comparing Information-Theoretic Measures of Complexity in Boltzmann Machines". In: *Entropy* 19.310 (2017). DOI: 10.3390/e19070310.

[Kim18] H. Kim, A. Hudetz, J. Lee, G. Mashour, and U. Lee. "Estimating the Integrated Information Measure Phi from High-Density Electroencephalography during States of Consciousness in Humans". In: *Frontiers in Human Neuroscience* 12 (2018). ISSN: 1662-5161. DOI: 10.3389/fnhum.2018.00042.

[Kinsner10] W. Kinsner. "System Complexity and Its Measures: How Complex Is Complex". In: *Advances in Cognitive Informatics and Cognitive Computing*. Ed. by Y. Wang, D. Zhang, and W. Kinsner. Berlin, Heidelberg: Springer, 2010, pp. 265–295. DOI: 10.1007/978-3-642-16083-7_14.

[Kirsh94] D. Kirsh and P. Maglio. "On Distinguishing Epistemic from Pragmatic Action". In: *Cognitive Science* 18.4 (1994), pp. 513–549. DOI: https://doi.org/10.1207/s15516709cog1804_1.

[Kiverstein09] J. Kiverstein and A. Clark. "Introduction: Mind Embodied, Embedded, Enacted: One Church or Many?" In: *Topoi* 28 (2009), pp. 1–7. DOI: 10.1007/s11245-008-9041-4.

[Kleiner21] J. Kleiner and S. Tull. "The Mathematical Structure of Integrated Information Theory". In: *Frontiers in Applied Mathematics and Statistics* 6 (June 2021). DOI: 10.3389/fams.2020.602973.

[Klyubin04] A. S. Klyubin, D. Polani, and C. L. Nehaniv. "Tracking Information Flow through the Environment: Simple Cases of Stigmergy". In: *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*. 2004, pp. 563–568. ISBN: 978-0-262-66183-6. URL: https://uhra.herts.ac.uk/bitstream/handle/2299/3945/101985.pdf?sequence=4&isAllowed=y (visited on 08/10/2023).

[Klyubin05] A. Klyubin, D. Polani, and C. Nehaniv. "Empowerment: a universal agent-centric measure of control". In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 1. 2005, pp. 128–135. DOI: 10.1109/CEC.2005.1554676.

[Klyubin07] A. S. Klyubin, D. Polani, and C. L. Nehaniv. "Representations of Space and Time in the Maximization of Information Flow in the Perception-Action Loop". In: *Neural computation* 19 (Oct. 2007), pp. 2387–432. DOI: 10.1162/neco.2007.19.9.2387.

[Koch16] C. Koch, M. Massimini, M. Boly, and G. Tononi. "Neural correlates of consciousness: Progress and problems". In: *Nature Reviews Neuroscience* 17 (2016), pp. 307–321. DOI: 10.1038/nrn.2016.22.

[Kolchinsky22] A. Kolchinsky. "A Novel Approach to the Partial Information Decomposition". In: *Entropy* 24.3 (2022), p. 403. ISSN: 1099-4300. DOI: 10.3390/e24030403.

[Kubjas15] K. Kubjas, E. Robeva, and B. Sturmfels. "Fixed points of the EM algorithm and nonnegative rank boundaries". In: *The Annals of Statistics* 43.1 (2015), pp. 422–461. DOI: 10.1214/14-AOS1282.

[Kück06] H. Kück. "Die Heider-Simmel-Studie (1944) in neueren Replikationen". In: *Gruppendynamik* 37 (2006), pp. 185–196. URL: 10.1007/s11612-006-0021-0.

[Kwok03] P. Kwok, W. Kong, K. Kasturi, C. Lee, and J. Hamill. "A Biomechanical Study on the Parachute Landing Fall". In: *17th AIAA Aerodynamic Decelerator Systems Technology Conference and Seminar*. 2003. ISBN: 978-1-62410-087-1. DOI: 10.2514/6.2003-2149.

[Ladyman13] J. Ladyman, J. Lambert, and K. Wiesner. "What is a complex system?" In: *European Journal for Philosophy of Science* 3 (2013), pp. 33–67. DOI: 10.1007/s13194-012-0056-8.

[Langer20a] C. Langer. *Integrated-Information-Measures GitHub Repository*. 2020. URL: https://github.com/CarlottaLanger/Integrated-Information-Measures (visited on 08/20/2022).

[Langer20b] C. Langer and N. Ay. "Complexity as Causal Information Integration". In: *Entropy* 22.10 (2020). DOI: 10.3390/e22101107.

[Langer21a] C. Langer. *Morphology Shapes Integrated Information GitHub Repository*. 2021. URL: https://github.com/CarlottaLanger/MorphologyShapesIntegratedInformation (visited on 07/20/2023).

[Langer21b] C. Langer and N. Ay. "How Morphological Computation Shapes Integrated Information in Embodied Agents". In: *Frontiers in Psychology* 12 (2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.716433.

[Langer22] C. Langer. *LearningRequiresIntInf GitHub Repository*. 2022. URL: https://github.com/CarlottaLanger/LearningRequiresIntInf (visited on 09/01/2022).

[Langer24] C. Langer and N. Ay. *Outsourcing Control requires Control Complexity*. Accepted at Artificial Life. 2024. DOI: https://doi.org/10.48550/arXiv.2209.01418.

[Lauritzen96] S. L. Lauritzen. *Graphical Models*. Oxford: Clarendon Press, 1996. ISBN: 978-0-191-59122-8.

[Lungarella05] M. Lungarella, T. Pegors, D. Bulwinkle, and O. Sporns. "Methods for quantifying the informational structure of sensory and motor data". In: *Neuroinformatics* 3.3 (2005), pp. 243–262. DOI: 10.1385/NI:3:3:243.

[Lungarella06] M. Lungarella and O. Sporns. "Mapping Information Flow in Sensorimotor Networks". In: *PLOS Computational Biology* 2.10 (Oct. 2006). DOI: 10.1371/journal.pcbi.0020144.

[MacKay03] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press, 2003. ISBN: 978-0-521-64298-9.

[Mallatt21] J. Mallatt. "A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness". In: *Entropy* 23.6 (2021), p. 650. DOI: 10.3390/e23060650.

[Maris96] M. Maris and R. Boeckhorst. "Exploiting physical constraints: heap formation through behavioral error in a group of robots". In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS '96*. Vol. 3. 1996, pp. 1655–1660. DOI: 10.1109/IROS.1996.569034.

[Marshall16] W. Marshall, J. Gomez-Ramirez, and G. Tononi. "Integrated Information and State Differentiation". In: *Frontiers in Psychology* 7 (2016). DOI: 10.3389/fpsyg.2016.00926.

[Marshall18] W. Marshall, L. Albantakis, and G. Tononi. "Black-boxing and cause-effect power". In: *PLOS Computational Biology* 14.4 (2018). DOI: 10.1371/journal.pcbi.1006114.

[Marshall22] W. Marshall, M. Grasso, W. G. Mayner, A. Zaeemzadeh, L. S. Barbosa, E. Chastain, G. Findlay, S. Sasai, L. Albantakis, and G. Tononi. *System Integrated Information*. 2022. DOI: 10.48550/ARXIV.2212.14537.

[Marstaller13] L. Marstaller, A. Hintze, and A. C. "The evolution of representation in simple cognitive networks". In: *Neural Computation* (2013), pp. 2079–2107. DOI: 10.1162/NECO_a_00475.

[Massimini05] M. Massimini, F. Ferrarelli, R. Huber, S. Esser, H. Singh, and G. Tononi. "Breakdown of Cortical Effective Connectivity During Sleep". In: *Science (New York, N.Y.)* 309 (Oct. 2005), pp. 2228–2232. DOI: 10.1126/science.1117256.

[Massimini10] M. Massimini, F. Ferrarelli, M. Murphy, R. Huber, B. Riedner, S. Casarotto, and G. Tononi. "Cortical reactivity and effective connectivity during REM sleep in humans". In: *Cognitive neuroscience* 1 (2010), pp. 176–183. DOI: 10.1080/17588921003731578.

[McGeer90a] T. McGeer. "Passive walking with knees". In: *Proceedings., IEEE International Conference on Robotics and Automation*. Vol. 3. 1990, pp. 1640–1645. DOI: 10.1109/ROBOT.1990.126245.

[McGeer90b] T. McGeer. "Passive Dynamic Walking". In: *The International Journal of Robotics Research* 9.2 (1990), pp. 62–82. DOI: 10.1177/027836499000900206.

[Mediano18] P. A. Mediano, A. K. Seth, and A. B. Barrett. "Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation". In: *Entropy* 21.1 (2018). ISSN: 1099-4300. DOI: 10.3390/e21010017.

[Mediano22a] P. A. M. Mediano, F. E. Rosas, J. C. Farah, M. Shanahan, D. Bor, and A. B. Barrett. "Integrated information as a common signature of dynamical and information-processing complexity". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32.1 (2022). DOI: 10.1063/5.0063384.

[Mediano22b] P. A. Mediano, F. E. Rosas, D. Bor, A. K. Seth, and A. B. Barrett. "The strength of weak integrated information theory". In: *Trends in Cognitive Sciences* 26.8 (2022), pp. 646–655. DOI: 10.1016/j.tics.2022.04.008.

[Merker22] B. Merker, K. Williford, and D. Rudrauf. "The integrated information theory of consciousness: A case of mistaken identity". In: *Behavioral and Brain Sciences* 45 (2022). DOI: 10.1017/S0140525X21000881.

[Miłkowski18] M. Miłkowski. "Morphological Computation: Nothing but Physical Computation". In: *Entropy* 20.12 (2018). DOI: 10.3390/e20120942.

[Montúfar12] G. Montúfar. "On the Expressive Power of Discrete Mixture Models, Restricted Boltzmann Machines, and Deep Belief Networks-A Unified Mathematical Treatment". PhD thesis. Universität Leipzig, 2012.

[Montúfar15] G. Montúfar, K. Ghazi-Zahedi, and N. Ay. "A Theory of Cheap Control in Embodied Systems". In: *PLOS Computational Biology* 11.9 (2015). DOI: 10.1371/journal.pcbi.1004427.

[Müller17] V. Müller and M. Hoffmann. "What Is Morphological Computation? On How the Body Contributes to Cognition and Control". In: *Artificial Life* 23 (2017). DOI: 10.1162/ARTL_a_00219.

[Nakajima13a] K. Nakajima, H. Hauser, R. Kang, E. Guglielmino, D. Caldwell, and R. Pfeifer. "A soft body as a reservoir: case studies in a dynamic model of octopus-inspired soft robotic arm". In: *Frontiers in Computational Neuroscience* 7 (2013). DOI: 10.3389/fncom.2013.00091.

[Nakajima13b] K. Nakajima, H. Hauser, R. Kang, E. Guglielmino, D. G. Caldwell, and R. Pfeifer. "Computing with a muscular-hydrostat system". In: *2013 IEEE International Conference on Robotics and Automation*. 2013, pp. 1504–1511. DOI: 10.1109/ICRA.2013.6630770.

[Nakajima14] K. Nakajima, T. Li, H. Hauser, and R. Pfeifer. "Exploiting short-term memory in soft body dynamics as a computational resource". In: *Journal of The Royal Society Interface* 11 (2014). DOI: 10.1098/rsif.2014.0437.

[Nguyen21] P. D. H. Nguyen, Y. K. Georgie, E. Kayhan, M. Eppe, V. V. Hafner, and S. Wermter. "Sensorimotor Representation Learning for an "Active Self" in Robots: A Model Survey". In: *KI-Künstliche Intelligenz* 35 (2021), pp. 9–35.

[Oizumi14] M. Oizumi, L. Albantakis, and G. Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0". In: *PLOS Computational Biology* (2014). DOI: 10.1371/journal.pcbi.1003588.

[Oizumi16a] M. Oizumi, S. Amari, T. Yanagawa, N. Fujii, and N. Tsuchiya. "Measuring Integrated Information from the Decoding Perspective". In: *PLOS Computational Biology* (2016). DOI: 10.1371/journal.pcbi.1004654.

[Oizumi16b] M. Oizumi, N. Tsuchiya, and S. Amari. "Unified framework for information integration based on information geometry". In: *JournalProceedings of the National Academy of Sciences of the United States of America*. 2016, pp. 14817–14822. DOI: 10.1073/pnas.1603583113.

[Papazoglou01] L. Papazoglou, A. Galatos, M. Patsikas, I. Savvas, L. Leontides, M. Trifonidou, and M. Karayannopoulou. "High-rise Syndorme in Cats: 207 cases (1988-1998)". In: *Australian Veterinary Practitioner* 31 (2001), pp. 98–102. ISSN: 0310-138X.

[Pautz19] A. Pautz. "What is the Integrated Information Theory of Consciousness? A Catalogue of Questions". In: *Journal of Consciousness Studies* (2019), pp. 188–215. URL: https://philarchive.org/archive/PAUWIT-2 (visited on 07/10/2023).

[Pearl09] J. Pearl. *Causality*. Cambridge: Cambridge University Press, 2009. ISBN: 978-0-511-80316-1.

[Pearl19] J. Pearl. "On the Interpretation of do(x)". In: *Journal of Causal Inference* 7 (2019). DOI: 10.1515/jci-2019-2002.

[Pearl85a] J. Pearl. *Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning*. Tech. rep. CSD-850021, R-43. UCLA Computer Science Department, 1985.

[Pearl85b] J. Pearl and A. Paz. *Graphoids: A Graph-Based Logic for Reasoning About Relevance Relations*. Tech. rep. CSD-850038, R-53. UCLA Computer Science Department, 1985.

[Pearl88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann Publishers, 1988. ISBN: 978-1-55860-479-7.

[Peretto84] P. Peretto. "Collective properties of neural networks: a statistical physics approach". In: *Biological Cybernetics* 50 (1984), pp. 51–62. DOI: 10.1007/BF00317939.

[Peyré19] G. Peyré and M. Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607. DOI: 10.1561/2200000073.

[Pfeifer01] R. Pfeifer and C. Scheier. *Understanding intelligence*. Cambridge: MIT Press, 2001. ISBN: 978-0-262-25679-7.

[Pfeifer03] R. Pfeifer and F. Iida. "Embodied Artificial Intelligence: Trends and Challenges". In: *Embodied Artif. Intell.* Vol. 3139. Jan. 2003, pp. 1–26. ISBN: 978-3-540-22484-6. DOI: 10.1007/978-3-540-27833-7_1.

[Pfeifer06a] R. Pfeifer and J. Bongard. *How the body shapes the way we think: a new view of intelligence*. Cambridge: The MIT Press (Bradford Books), 2006. ISBN: 978-0-262-53742-1.

[Pfeifer06b] R. Pfeifer, F. Iida, and G. Gomez. "Morphological computation for adaptive behavior and cognition". In: *International Congress Series* 1291 (June 2006), pp. 22–29. DOI: 10.1016/j.ics.2005.12.080.

[Pfeifer07] R. Pfeifer, M. Lungarella, and F. Iida. "Self-Organization, Embodiment, and Biologically Inspired Robotics". In: *Science* 318.5853 (2007), pp. 1088–1093. DOI: 10.1126/science.1145803.

[Pfeifer09] R. Pfeifer and G. Gómez. "Morphological Computation – Connecting Brain, Body, and Environment". In: *Creating Brain-Like Intelligence: From Basic Principles to Complex Intelligent Systems*. Ed. by B. Sendhoff, E. Körner, O. Sporns, H. Ritter, and K. Doya. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 66–83. DOI: 10.1007/978-3-642-00616-6_5.

[Pfeifer98] R. Pfeifer and C. Scheier. "Representation in Natural and Artificial Agents: An Embodied Cognitive Science Perspective". In: *Zeitschrift für Naturforschung C* 53.7-8 (1998), pp. 480–503. DOI: doi:10.1515/znc-1998-7-804.

[Pietra95] S. D. Pietra, V. D. Pietra, and J. Lafferty. "Inducing features of random fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.4 (1995), pp. 380–393. DOI: 10.1109/34.588021.

[Polani07] D. Polani, O. Sporns, and M. Lungarella. "How Information and Embodiment Shape Intelligent Information Processing". In: *50 Years of Artificial Intelligence*. Ed. by M. Lungarella, F. Iida, J. Bongard, and R. Pfeifer. Vol. 4850. Berlin, Heidelberg: Springer, 2007, pp. 99–111. DOI: 10.1007/978-3-540-77296-5_10.

[Polani09] D. Polani and M. Möller. "Models of Information Processing in the Sensorimotor Loop". In: *Information Theory and Statistical Learning*. Ed. by F. Emmert-Streib and M. Dehmer. Boston: Springer US, 2009, pp. 289–308. DOI: 10.1007/978-0-387-84816-7_12.

[Powers73] W. T. Powers. *Behavior: The Control Of Perception*. New Canaan, Connecticut: Benchmark Publications Inc., 1973. ISBN: 0-9647121-7-2.

[Puddle13] D. Puddle and P. Maulder. "Ground Reaction Forces and Loading Rates Associated with Parkour and Traditional Drop Landing Techniques". In: *Journal of sports science & medicine* 12 (Mar. 2013), pp. 122–129.

[Rauh13] J. Rauh. "Optimally approximating exponential families". eng. In: *Kybernetika* 49.2 (2013), pp. 199–215. URL: http://eudml.org/doc/260703 (visited on 02/09/2022).

[Sadeghi16] K. Sadeghi. "Marginalization and conditioning for LWF chain graphs". In: *The Annals of Statistics* (2016). DOI: 10.1214/16-aos1451.

[Salge14] C. Salge, C. Glackin, and D. Polani. "Empowerment–An Introduction". In: *Guided Self-Organization: Inception*. Ed. by M. Prokopenko. Berlin, Heidelberg: Springer, 2014, pp. 67–114. DOI: 10.1007/978-3-642-53734-9_4.

[Seitzer21] M. Seitzer, B. Schölkopf, and G. Martius. "Causal Influence Detection for Improving Efficiency in Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. 2021. URL: https://openreview.net/forum?id=DXJl9826dm (visited on 07/10/2023).

[Shannon48] C. E. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27 (1948), pp. 379–423. ISSN: 0005-8580.

[Sporns04] O. Sporns and T. K. Pegors. "Information-Theoretical Aspects of Embodied Artificial Intelligence". In: *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*. Ed. by F. Iida, R. Pfeifer, L. Steels, and Y. Kuniyoshi. Berlin, Heidelberg: Springer, 2004, pp. 74–85. ISBN: 978-3-540-22484-6. DOI: 10.1007/978-3-540-27833-7_5.

[Stewart10] J. Stewart. "Foundational Issues in Enaction as a Paradigm for Cognitive Science: From the Origin of Life to Consciousness and Writing". In: *Enaction: Toward a New Paradigm for Cognitive Science* (Nov. 2010), pp. 1–32. DOI: 10.7551/mitpress/9780262014601.003.0002.

[Studený05] M. Studený. *Probabilistic Conditional Independence Structures*. London: Springer, 2005. ISBN: 978-1-852-33891-6.

[Studený89] M. Studený. "Multiinformation and the Problem of Characterization of Conditional Independence Relations". In: *Problems of Control and Information Theory* 18 (1989), pp. 3–16.

[Studený97] M. Studený. "Semigraphoids and structures of probabilistic conditional independence". In: *Annals of Mathematis and Artificial Intelligence* (1997), pp. 71–98. DOI: 10.1023/A:1018905100242.

[Sullivant07] S. Sullivant. "Toric fiber products". In: *Journal of Algebra*. Computational Algebra 316.2 (2007), pp. 560–577. ISSN: 0021-8693. DOI: 10.1016/j.jalgebra.2006.10.004.

[Sullivant18] S. Sullivant. *Algebraic Statistics*. Providence: American Mathematical Society, 2018. ISBN: 978-1-470-43517-2.

[Thelen94] E. Thelen and L. Smith. *A dynamic systems approach to the development of cognition and action*. Cambridge: The MIT Press, 1994. ISBN: 978-0-262-28487-5.

[Tishby11] N. Tishby and D. Polani. "Information Theory of Decisions and Actions". In: *Perception-Action Cycle: Models, Architectures, and Hardware*. Ed. by V. Cutsuridis, A. Hussain, and J. G. Taylor. New York: Springer, 2011, pp. 601–636. DOI: 10.1007/978-1-4419-1452-1_19.

[Tononi03] G. Tononi and O. Sporns. "Measuring information integration". In: *BMC Neuroscience* 4 (2003). DOI: https://doi.org/10.1186/1471-2202-4-31.

[Tononi04] G. Tononi. "An information integration theory of consciousness". In: *BMC Neuroscience* 5 (2004). DOI: 10.1186/1471-2202-5-42.

[Tononi08] G. Tononi. "Consciousness as Integrated Information: a Provisional Manifesto". In: *The Biological Bulletin* 215 (2008), pp. 216–242. DOI: 10.2307/25470707.

[Tononi12] G. Tononi. "Integrated information theory of consciousness: An updated account". In: *Archives italiennes de biologie* 150 (June 2012), pp. 56–90. DOI: 10.4449/aib.v149i5.1388.

[Tononi16a] G. Tononi, M. Boly, O. Gosseries, and S. Laureys. "The Neurology of Consciousness". In: Dec. 2016, pp. 407–461. DOI: 10.1016/B978-0-12-800948-2.00025-X.

[Tononi16b] G. Tononi, M. Boly, M. Massimini, and C. Koch. "Integrated information theory: From consciousness to its physical substrate". In: *Nature Reviews Neuroscience* 17 (May 2016), pp. 450–461. DOI: 10.1038/nrn.2016.44.

[Tononi94] G. Tononi, O. Sporns, and G. M. Edelman. "A measure for brain complexity: Relating functional segregation and integration in the nervous system". In: *Proceedings of the National Academy of Sciences of the United States of America* (1994), pp. 5033–5037. DOI: 10.1073/pnas.91.11.503.

[Tononi98a] G. Tononi and G. M. Edelman. "Consciousness and Complexity". In: *Science* 282 (1998), pp. 1846–1851. DOI: 10.1126/science.282.5395.1846.

[Tononi98b] G. Tononi, G. M. Edelman, and O. Sporns. "Complexity and coherency: integrating information in the brain". In: *Trends in Cognitive Sciences* (1998), pp. 474–484. ISSN: 1364-6613. DOI: 10.1016/S1364-6613(98)01259-5.

[Touchette04] H. Touchette and S. Lloyd. "Information-theoretic approach to the study of control systems". In: *Physica A: Statistical Mechanics and its Applications* 331.1 (2004), pp. 140–172. DOI: 10.1016/j.physa.2003.09.007.

[Toussaint06] M. Toussaint, S. Harmeling, and A. Storkey. *Probabilistic inference for solving (PO)MDPs*. Tech. rep. 934. School of Informatics, University of Edinburgh, Dec. 2006.

[Toussaint08] M. Toussaint, L. Charlin, and P. Poupart. "Hierarchical POMDP Controller Optimization by Likelihood Maximization". In: *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence*. 2008, pp. 562–570. DOI: 10.5555/3023476.3023543.

[Toussaint09] M. Toussaint. "Probabilistic inference as a model of planned behavior". In: *Künstliche Intelligenz* 23.3 (2009), pp. 23–29.

[Uexküll92] J. Von Uexküll. "A stroll through the worlds of animals and men: A picture book of invisible worlds". In: *Semiotica* 89.4 (1992), pp. 319–391. DOI: doi:10.1515/semi.1992.89.4.319.

[Varela91] F. J. Varela, E. Rosch, and E. Thompson. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: The MIT Press, Sept. 1991. ISBN: 978-0-262-28547-6. DOI: 10.7551/mitpress/6730.001.0001.

[Virgo19] N. Virgo. Personal communication on the design of the racetrack and agents. 2019.

[Vnuk04] D. Vnuk, B. Pirkić, D. Matičić, B. Radišić, M. Stejskal, T. Babić, M. Kreszinger, and N. Lemo. "Feline high-rise syndrome: 119 cases (1998–2001)". In: *Journal of Feline Medicine and Surgery* 6.5 (2004), pp. 305–312. DOI: 10.1016/j.jfms.2003.07.001.

[Vomlel99] J. Vomlel. "Methods of Probabilistic Knowledge Integration". Dissertation. Czech Technical University, Faculty of Electrical Engineering, 1999. URL: http://staff.utia.cas.cz/vomlel/prace.pdf (visited on 02/09/2022).

[Williams10] P. L. Williams and R. D. Beer. *Nonnegative Decomposition of Multivariate Information*. 2010. DOI: 10.48550/arXiv.1004.25155.

[Wilson02] M. Wilson. "Six views of embodied cognition". In: *Psychonomic Bulletin & Review* 9 (2002), pp. 625–636. DOI: https://doi.org/10.3758/BF03196322.

[Winkler03] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer, 2003. ISBN: 978-3-540-44213-4.

[Wolpert95a] D. M. Wolpert and Z. Ghahramani. "Computational principles of movement neuroscience". In: *Nature Neuroscience* (1995), pp. 1212–1217. DOI: 10.1038/81497.

[Wolpert95b] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan. "An Internal Model for Sensorimotor Integration". In: *Science* 269.5232 (1995), pp. 1880–1882. DOI: 10.1126/science.7569931.

[Zanardi18] P. Zanardi, M. Tomka, and L. C. Venuti. *Towards Quantum Integrated Information Theory*. 2018. DOI: 10.48550/ARXIV.1806.01421.

[Zhang14a] Z. Zhang, J. Yang, and H. Yu. "Effect of Flexible Back on Energy Absorption during Landing in Cats: A Biomechanical Investigation". In: *Journal of Bionic Engineering* 11.4 (2014), pp. 506–516. ISSN: 1672-6529. DOI: 10.1016/S1672-6529(14)60063-9.

[Zhang14b] Z. Zhang, H. Yu, J. Yang, L. Wang, and L. Yang. "How cat lands: Insights into contribution of the forelimbs and hindlimbs to attenuating impact force". In: *Chinese Science Bulletin* 59 (2014), pp. 3325–3332. DOI: 10.1007/s11434-014-0328-0.

Index

List of Figures