INFORMATION DRIVEN SELF-ORGANIZATION OF AGENTS AND AGENT COLLECTIVES

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

Malte Harder

May 2013

Malte Harder
Ansbacher Str. 69a
28215 Bremen
Germany
me@malteharder.de

Typeset in Cardo v1.04 and Adobe Source Sans Pro v1.038 with ConTEXt, TikZ, and PGFPlots.
Layout inspired by *The Elements of Typographic Style* by Robert Bringhurst.
Comics from http://www.xkcd.com (licensed under CC-BY-NC 2.5).

Typeset on Thursday April 3, 2014.

ABSTRACT

From a visual standpoint it is often easy to point out whether a system is considered to be self-organizing or not, though a quantitative approach would be more helpful. Information theory, as introduced by Shannon, provides the right tools not only quantify self-organization, but also to investigate it in relation to the information processing performed by individual agents within a collective.

This thesis sets out to introduce methods to quantify spatial self-organization in collective systems in the continuous domain as a means to investigate morphogenetic processes. In biology, morphogenesis denotes the development of shapes and form, for example embryos, organs or limbs. Here, I will introduce methods to quantitatively investigate shape formation in stochastic particle systems.

In living organisms, self-organization, like the development of an embryo, is a guided process, predetermined by the genetic code, but executed in an autonomous decentralized fashion. Information is processed by the individual agents (e.g. cells) engaged in this process. Hence, information theory can be deployed to study such processes and connect self-organization and information processing. The existing concepts of observer based self-organization and relevant information will be used to devise a framework for the investigation of guided spatial self-organization.

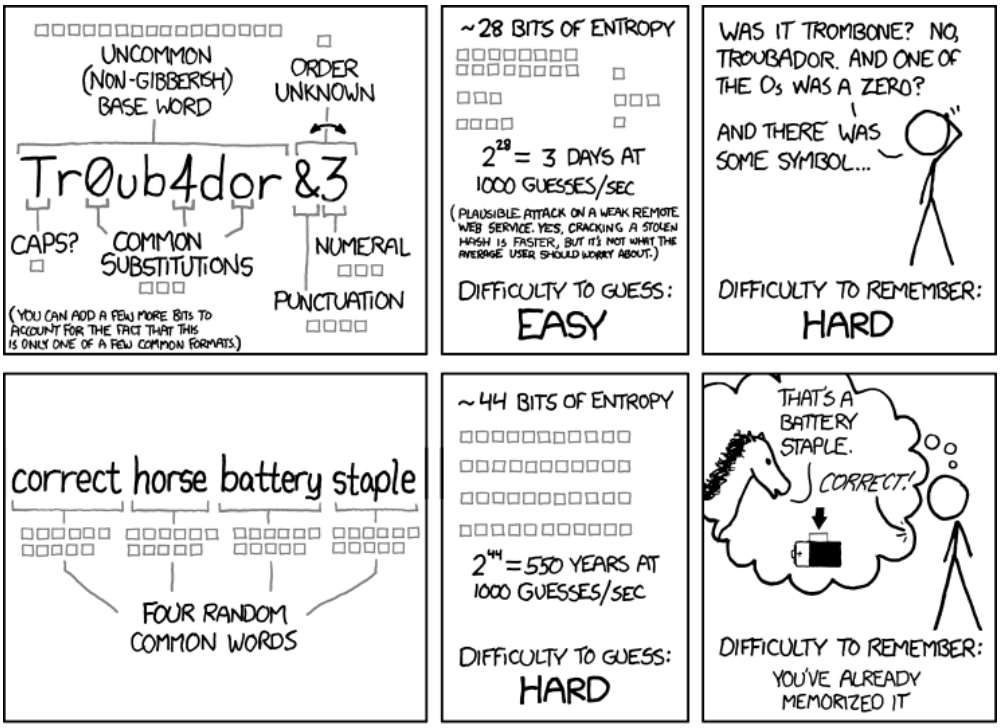
Furthermore, local information transfer plays an important role for processes of self-organization. In this context, the concept of synergy has been getting a lot attention lately. Synergy is a formalization of the idea that for some systems the whole is more than the sum of its parts and it is assumed that it plays an important role in self-organization, learning and decision making processes. In this thesis, a novel measure of synergy will be introduced, that addresses some of the theoretical problems that earlier approaches posed.

» *The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny...'* «

ISAAC ASIMOV, Unknown

» *The difference between life and non-life is a matter not of substance but of information.* «

RICHARD DAWKINS, The Greatest Show on Earth



» *To anyone who understands information theory and security and is in an infuriating argument with someone who does not (possibly involving mixed case), I sincerely apologize* «

XKCD, 936

ACKNOWLEDGEMENTS

I thank my principal supervisor Dr. Daniel Polani and my secondary supervisor Prof. Chrystopher Nehaniv for many inspiring and invaluable discussions which inarguably had a significant impact on my research. I am also thanking everybody who provided feedback to this thesis and the articles that I published during this programme.

Furthermore, I am grateful to my former house mates from Gloucester Court for the many cups of tea we had, that were often accompanied by scientific discussions, to my friends for offering the occasional escape from information theory, to my family for always supporting me, especially during the last months of writing this thesis. Special thanks go to my brother for keeping me from going insane on several occasions by 'showing me where that one semicolon was missing'.

CONTENTS

1 INTRODUCTION 15
Motivation 15
Overview 17
Contribution 18
2 BACKGROUND 20
Information Theory 20
Related Work 28
3 QUANTIFYING SELF-ORGANIZATION 37
Introduction 37
Self-Organization & Complexity 37
Statistical Complexity 41
Self-organization via Observers 46
Comparison of SC-Organization and O-Organization 51
Estimation of Multi-information 53
Discussion 62
4 SELF-ORGANIZATION OF PARTICLE SYSTEMS 68
Particle Collectives & Self-organisation 69
Methods 74
Examples 76
Results 79
Discussion 88
5 INFOGENESIS 90
Introduction 90
Information-theoretic Control Theory 91
Relevant Information 93
Embodiment & Perception-action Loops 96
Multi-Agent Relevant Information 99
Relevant Information & Self-Organization 102
Episodic Tasks & Shapes as Goals 103
Shared Control and Sensor Coordination 106
Experiments 108
Discussion 111
6 REDUNDANT INFORMATION 114
What is Redundancy 114
Measure Candidates 115
Construction of a New Measure 119
Partial Information Decomposition 125

Comparisons 135
Mechanistic & Source Redundancy 141
Information Transfer 142
Multivariate Extensions 150
Discussion 153
7 CONCLUSION 156
Summary 156
Discussion 157
Future Work 160
BIBLIOGRAPHY 162

TABLE OF FIGURES

2.1 Illustration of relations between (conditional) entropy and (conditional) mutual information. 24
2.2 Illustration of the CBN of the perception-action loop of a memoryless agent. 27
2.3 Illustration of the relevant information trade-off curve. To reach a higher utility more information needs to be processed. The top right corner of the curve marks the relevant information, the amount of information that is needed to follow a policy that is optimal in utility. The shape of the curve in this illustration is typical for relevant information curves of reinforcement learning scenarios: Much of the information needed by the agent only contributes to a small amount of utility at the top of the curve. 30
2.4 Examples of Turing patterns generated with diffusion-reaction systems. 33
3.1 The "one-humped complexity curve" as introduced by Crutchfield and Young (1989), illustrating the idea that systems of high complexity lie somewhere between simple deterministic and completely random systems. 38
3.2 Illustration of a (universal) Turing machine (based on an illustration from http://texample.net by Ludger Humbert licensed under CC-BY 2.5). A Turing machine consists of a finite control (finite number of states) a reading and writing head and an infinite tape consisting of symbols of an alphabet and a set of transition rules from state and input to a new state and a tapte shift (Turing, 1936). A Universal Turing Machine is a Turing Machine that first reads in the description of a Turing Machine and then simulates this Turing Machine on arbitrary input. 41
3.3 Illustration of a Bernoulli Turing Machine (based on an illustration from http://texample.net by Ludger Humbert licensed under CC-BY 2.5). A Bernoulli Turing Machine is a Universel Turing Machine which transition rules have access to a register containing true (thermodynamical) randomness. 42
3.4 Illustration of systems exhibiting different amounts of multi-information. Each row depicts samples from different distributions of points in the unit square. The first two rows illustrate the extreme cases of low multi-information (between the random variables of the positions of individual points), in the top row there is no variation at all and in the second row there is no correlation between the points. The bottom rows show samples from distributions with a high degree of correlation, which therefore exhibit a larger amount of multi-information. 47
3.5 Showing the dependencies of random variables at different times for the observer choice example. Connected variables are copies of each other, unconnected variabled are independent. 48
3.6 Comparison of different binning estimators and binning rules on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate) (part 1/2, see Figure 3.7 for part 2/2). Error bars denote one standard deviation from 50 estimation samples. 63
3.7 Comparison of different binning estimators and binning rules on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate) (part 2/2, see Figure 3.6 for part 1/2). Error bars denote one standard deviation from 50 estimation samples. 64

3.8 Comparison of different values of k (the k-th neighbour is used in the algorithm to estimate the multi-information) for the **KSG** estimator on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate) (part 2/2, see Figure 3.6 for part 1/2). Error bars denote one standard deviation from 50 estimation samples. 65
3.9 Comparison of **KSG** estimator and kernel density estimation on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate). The plot for $d = 50$ and the **UBX** scenario is missing as the analytical calculation of the mutual information exceeds memory limits in this case. Error bars denote one standard deviation from 50 estimation samples. 66
3.10 Comparison of estimates with different sample sizes ($m = 100, 500, 2500$) in high dimensional settings ($d = 100, 200$) using the **KSG** estimator. Error bars denote one standard deviation from 50 estimation samples. 67
4.1 Example of a particle configuration. 69
4.2 Plot of the force scaling function used for the particle dynamics, $r_{\alpha\beta}$ denotes the prefered distance between particles of type α and β. This radius can be directly specified as a parameter of the function. The long range attraction of is cut off by the radius r_c, whereas r_b limits the repellent force for particle that are very close. 71
4.3 Examples of equilibrium states of particle collectives with different number of types. 71
4.4 Multi-information between three particles of the same type for different noise levels ($m = 1000$ samples, $r_c = 10$, $r_{\alpha\alpha} = 2.5$). 76
4.5 Plot of the samples from the noise free three particle example at two different time steps. The particle configurations are shaded by sample. Therefore, it can be seen in a) that the outliers along the three axes belong to the same samples. This is a sign for correlation between the particles and hints towards a larger amount of multi-information. 77
4.6 Multi-information between three particles with $l = 1$ and $l = 3$ types ($m = 1000$ samples, $r_c = 10$, $r_{\alpha\beta} = [[2.5, 0.1, 5], [0.1, 1, 0.5], [5, 0.5, 2]]$ and $k_{\alpha\beta} = [[0.05, 0.5, 0.1], [0.5, 0.2, 0.5], [0.1, 0.5, 0.3]]$). 78
4.7 Plot of all samples of particle configurations of the three particle and three types example at different time steps. Each shade denotes a different type, all samples of configurations of three particles of three different types are overlayed in this plot. 80
4.8 Multi-information between particles plotted against time with $n = 70$, $l = 3$, $r_c = 6.0$, $r_{\alpha\beta} = [[2.5, 5, 4], [5, 2.5, 2], [4, 2, 3.5]]$ and $k_{\alpha\beta} = [[0.6, 0.1, 0.1], [0.1, 0.6, 0.6], [0.1, 0.6, 0.6]]$ ($m = 500$ samples). The increase of multi-information correlates with the visual organization shown by snapshots of two samples at different times. 81
4.9 Increase of multi-information between $t = 0$ and $t = 5000$ for particle systems of different size ($m = 500$ samples, $r_c = 10$, $l = 1$ and $l = 3$ types using the same type specifications as in Figure 4.4 and Figure 4.6). 82
4.10 Plot of the samples from the noise free three particle example at two different time steps. The particle configurations are shaded by sample. Therefore, it can be seen in a) that the outliers along the three axes belong to the same samples. This is a sign for correlation between the particles and hints towards a larger amount of multi-information. 83

4.11 Multi-information between $n = 3$ and $n = 20$ particles of the same type with a smaller cut-off radius ($m = 1000$ samples, $r_c = 3$, $r_{\alpha\alpha} = 2.5$). 83
4.12 Plot of all particles of all samples at time $t = 5000$, the system consists of 20 particles of a single type, shading of the particle denotes different samples. In a) a regular grid can be seen and in b) it can be seen that the outer ring has been much better aligned so that for each particle samples match more closely (denser clusters), while this is not possible for the inner ring of particles as their alignment related to the outer ring is a degree of freedom. 85
4.13 Increase of multi-information between $t = 0$ and $t = 1500$, for different numbers of types ($n = 20$ particles, $r_c = 7.5$ and $m = 250$ samples). Averaged over 30 randomly generated types with mutual preferred distance radii $r_{\alpha\beta} \in [1.0, 5.0]$ and $k_{\alpha\beta} \in [0.25, 0.75]$. 86
4.14 Increase of multi-information between $t = 0$ and $t = 1500$, for different cut of radii r_c and numbers of types l, ($n = 20$ particles and $m = 250$ samples). Averaged over 30 randomly generated types with mutual preferred distance radii $r_{\alpha\beta} \in [1.0, 5.0]$ and $k_{\alpha\beta} \in [0.25, 0.75]$. 87
4.15 Contribution of the different terms of the decomposition normalized with the multi-information in each time-step. The total multi-information is normalized to fit the scale. 88
4.16 Examples of emergent structures in particle collectives. 89
5.1 CBN of control systems, W is the random variable that denotes the system state, C the controller and W' the system after control was applied. 91
5.2 Illustration of the CBN of the perception-action loop of a memoryless agent with full access to the world state. 94
5.3 Illustration of the relevant information formalism for a simple goal finding task in a 5×5 grid world. In a) the trade-off curve between relevant information and performance is shown, b) shows an optimal policy that can be the result of an optimization without any information constraint and c) - d) show policies for different value of β. 95
5.4 Illustration of the CBN of the perception-action loop of a memoryless agent. 97
5.5 Illustration of the CBN of the perception-action loop of an agent with memory. 99
5.6 Illustration of the CBNs of the perception-action loops of a collective of n memoryless agents. 99
5.7 Illustration of the CBNs of the perception-action loops of a collective of n memoryless agents. 101
5.8 Illustration of the CBN of the perception-action loop of a memoryless agent with full access to the world state. 104
5.9 In this 6×5 grid-world, the two dark-grey rectangles show the goal configuration, the light-grey rectangles show a configuration where the agents block each other if they move in the directions of the arrows. This causes that the agents stay at their current position. 108
5.10 Performance of agents with shared controller and individual controllers with summed expectation of utility per agent and relevant information for the joint distribution of $(a^{(1)}, a^{(2)})$. Both graphs show the same features but the scales differ. 109

5.11 Coordination of agents with shared controller on a 6×1 field, comparison of intrinsic coordination for shared control $I(\bar{A}^{(1)}; \bar{A}^{(2)}|W)$ with coordination for shared control $I(\bar{A}^{(1)}; \bar{A}^{(2)})$ and individual control $I(A^{(1)}; A^{(2)})$. 110
5.12 Coordination of agents with shared controllers in worlds of different sizes. 111
6.1 Illustration of the construction of projective information for binary input variables. Points represent the distributions in the space of distributions over the variable Z. The lines connecting points denote the subspace of conditional distributions depending on the distribution of X_1 and X_2 respectively. 121
6.2 Illustration of the Partial Information Decomposition into redundant, unique and synergistic terms. 127
6.3 Redundancy lattices for different sizes of index sets. Vertices represent elements of $\mathcal{A}(\mathbf{V})$, edges are connected if an element is ``smaller" with respect to the partial order $\preccurlyeq$ and there is no other element in $\mathcal{A}(\mathbf{V})$ that is smaller than the larger and larger than the smaller element. 129
6.4 Copy Example. Complete redundancy and complete uniqueness using I_{red}. 132
6.5 Comparison of total mutual information $I(Z; X_1, X_2)$ (dotted gray line), the new redundancy measure I_{red} (solid line) and I_{min} (dashed line) for varying values of λ, where λ controls the correllation between X_1 and X_2. It can be seen I_{min} measures a constant amount of redundancy and therefore does not distinguish between redundancy and uniqueness with varying λ as desired, whereas I_{red} does. 132
6.6 XOR Example. A purely synergistic mechanism. 133
6.7 AND Example. The total mutual information is $I(Z; X_1, X_2) = 0.811278$. 134
6.8 Plot of the redundant information $I_{\text{red}}(R; D_1, D_2)$ depending on the correlation λ between the two dice D_1 and D_2. From top to bottom the summation coefficient is $\alpha = 1, ..., 6$. It can be seen that for independent dice $\lambda = 1$ the amount of redundancy depends on the mechanism that is used to sum the results, whereas on the other extreme, all redundancy comes from the correlation of the sources. 135
6.9 Comparison of I_{min} and I_{red} for randomly drawn distributions $p(x, y, z)$ with $|\mathcal{X}| = |\mathcal{Y}| = 3$ fixed sized sets, plotted for different sizes of $\mathcal{Z}$. The change of $|\mathcal{Z}|$ also changes the dimension of the simplex where the distributions P_Z are contained in. Note that as the dimension of $\mathcal{Z}$ goes up, I_{min} gets larger in comparison to I_{red}. The distributions were drawn using a uniform distribution on a random subsimplex of $\Delta(X, Y, Z)$. The subsimplex was selected in each draw randomly with the probability of $p(x, y, z) = 0$ being 0.5 for each triple (x, y, z). 136
6.10 Conditional distributions visualized in $\Delta(Z)$ on the unit interval. 139
6.11 Illustration of the construction of projective information for binary input variables. The illustration shows why left monotonicity does not hold for I_{red}. 140
6.12 PI-diagram for the decomposition of transfer entropy into PI-atoms. The coloured areas denote the transfer entropy. 143
6.13 Bayesian network of the first example process. If $x_t = 0$ then x_{t+1} is a copy of y_t, if $x_t = 1$ then the bit of x_{t+1} is a flipped copy y_t. The probability that the bit is flipped in the copy is denoted by d. 143

6.14 Decomposition of transfer entropy $T_{Y \to X}$ for the first example process. The plot shows SITE (solid black line using I_{red}, dashed black line using I_{min}) and SDTE (solid gray line using I_{red}, dashed gray line using I_{min}) given d. It can be seen that both decompositions coincide for this process. 144
6.15 Bayesian network of the second example process. X_t is a parallel and independent process, the only information transfer between the processes is from Z_t to Y_{t+1}. 145
6.16 Decomposition of transfer entropy $T_{Z \to (X,Y)}$ for the second example process. The plot shows SITE (solid black line using I_{red}, dashed black line using I_{min}) and SDTE (solid gray line using I_{red}, dashed gray line using I_{min}). 146
6.17 Decomposition of transfer entropy $T_{Z \to Y}$ for the second example process. The plot shows SITE (dashed black line using I_{min}), SDTE dashed gray line using I_{min}). 146
6.18 The plot shows $I(Y_{t+1}; Z_t)$ (dotted gray line) and $I_{\text{red}}(Y_{t+1}; Y_t, Z_t)$ (solid black line) for the second example process. 147
6.19 Illustration of the conditional distributions of Y_{t+1} for the second example process in the two cases $d \leq 0.5$ and $d \geq 0.5$. The line represents the one dimensional simplex, i.e. the space of probability distributions over Y_{t+1} denoted by $\Delta(Y_{t+1})$ where Y_{t+1} is a binary valued random variable. The black diamond represents the marginal distribution of $p(y_{t+1})$ and the shaded diamonds the conditionals given specific values of Y_t and Z_t. It can now be seen that the projections are always equal to the conditional distributions closer to the marginal of Y_{t+1}. In particular, the projections are the same, no matter in which direction the projection is done (from Y_t to Z_t or vice versa). 148

INTRODUCTION

I

» *It is the pervading law of all things organic and inorganic,*
Of all things physical and metaphysical,
Of all things human and all things super-human,
Of all true manifestations of the head,
Of the heart, of the soul,
That the life is recognizable in its expression,
That form ever follows function. This is the law «

LOUIS SULLIVAN, The Tall Office Building Artistically Considered

1.1 MOTIVATION

Morphogenesis is the term used in biology to denote the development of shapes in organisms. One of the first theoretically inspired studies on morphogenesis were made by Alan Turing in the fifties with reaction-diffusion systems (Turing, 1952). Reaction diffusion systems are models of the concentration of substrates distributed in space. In these models two processes can change the concentration: local 'chemical' reactions and diffusion. This leads to the development of spatial patterns, like stripes, spots or spirals and explains how certain patterns can form from a homogeneous initial state (Harrison, 1994).

The study of morphogenesis in biology was pioneered by Thompson and Bonner (1992), but it was not until the development of modern genetics, molecular biology and importantly the discovery of DNA that the field gained traction. One of the earlier treatments of morphogenesis by Townes and Holtfreter (1955) considers the early stage of biological development in an organisms life, also known as embryogenesis, where a transformation from a simple ball of cells towards predetermined cell arrangements takes place. In particular, they studied the significance of cell-adhesion and cell-motility for the process of cell segregation and cell differentiation in the early stages of embryogenesis. Some of their experiments included the reshuffling of cells of different types that show selective cell-adhesion properties with respect to the cell type. In these cell aggregates they were able to show that over time cells segregated by type. These experiments were later modelled in simulations by Glazier and Graner (1993) using a model from statistical mechanics.

Subsequently, morphogenesis and especially embryogenesis stay very active fields of research in biology. Nowadays, there are ten known basic cellular mechanisms that seem to drive all morphogenetic processes in nature (Davies, 2005,2008). While the mechanisms of morphogenesis on a subcellular and molecular level are better explored today, less is known about interplay of these mechanism in the later stages of development (Davies, 2008). As the development of an organism is orchestrated by gene regulatory networks (Wolpert et al., 2002), the better understanding of these has an important impact on the study of morphogenesis. While functional cell differentiation or the development is more and more understood, the question about the principles guiding the development of a nervous system or the organization of muscles together with bones, tendons and ligaments remains

unsolved (Bard, 2008). Furthermore, are there a few common principles underlying all these processes?

The history of science has many examples where common principles were found via abstraction. In a very abstract sense, morphogenesis is a process of spatial self-organization guided by predetermined parameters that evolve over time. In this thesis, I want to propose an information-theoretic perspective on morphogenetic processes by connecting information-theoretic formulations of self-organization with methods from spatial statistics and information-theoretic models of multi-agent systems.

Information theory has come a long way from its initial purpose as a theory of communication by Shannon (1948). Here I will hold a view in the spirit of Ashby (1956), Barlow (1959) and Lwoff (1962) considering living organisms and, further down the hierarchy, their organs and their cells as information processing entities. Information, as defined by Shannon (1948), is a versatile measure that can be used to quantify the costs associated to decision making (Tishby and Polani, 2010) and requirements to sensor input with respect to extrinsic reward (Polani et al., 2006). It is versatile as it can be applied any model that can be captured in the language of random variables. Furthermore, it can be used to postulate conservation laws, bandwidth limitations or efficiency constraints. For example, information theory provides conservation laws concerning the information that needs to be injected into a system to control it (Touchette and Lloyd, 2004). This is of great relevance as conservation laws make it possible to state constraints and requirements that then can be applied to self-organising systems to make characterizations or predictions.

As information processing is usually associated with some kind of metabolic cost, information-theoretic constraints are grounded in the physical world. This opens an evolutionary perspective on information processing, where organisms that are parsimonious with information have an evolutionary advantage (Polani, 2009). In this context *parsimony* means that only information that is needed to achieve a certain level of fitness or adaptability is processed, and thus called *relevant*.

Self-organization, seen as an increase of complexity over time (Shalizi, 2001), can be the result of distributed information processing in collectives. This observation suggests that complexity as well can be captured by information-theoretic means and there are indeed several approaches using statistical and information-theoretic methods to define complexity and self-organization (Crutchfield and Young, 1989, Shalizi, 2001 and Polani, 2008).

This thesis can be divided into two parts. The first one deals with morphogenesis and shape formation of multi-agent systems, while the second part is about information-theoretic notions of synergy and redundancy. Synergy is the formalization of the idea that the whole is more than the sum of its parts. Though a direct relation to spatial self-organization is not yet discernible, there is much evidence, that synergy is an important driver of self-organization (Flecker et al., 2011 and Lizier et al., 2013).

In the main part of the thesis, I will show that currently available information-theoretic methods can be used to quantify self-organization in large collective systems. In particular, this method will be employed to study spatial stochastic dynamical systems that are modelled to roughly emulate biological cells. Finally, I will relate the information-theoretic notion of self organization to the concept of relevant information and provide a general information-theoretic framework to further investigate the constraints and properties of information processing in multi-agent that self-organize in a guided way towards spatial configurations.

1.2 OVERVIEW

This thesis is structured as follows:

Chapter 2 provides an introduction to the basics of information theory and Causal Bayesian Networks. Moreover, related literature is discussed with the aim to give an overview of applications of information theory to embodied cognition as well as self-organization.

Chapter 3 gives an introduction to the quantification of self-organization. Two measures, statistical complexity self-organization and observer self-organization are discussed regarding an application to spatial continuous systems. This chapter concludes with a quantitative comparison of multi-information estimation, which is used in the calculation of observer based self-organization.

Chapter 4 introduces a model of particle systems roughly mimicking biological cell motility and adhesion. A method to measure observer based self-organization from particle system simulations is developed. This method is then used to investigate the influence of types and particle interactions on the system's self-organization.

Chapter 5 formally introduces the concept of relevant information for multi-agent systems and discusses the perception-action loop as a model of embodied cognition, including the control theoretic implications of such a model. It is then shown how morphogenetic tasks can be investigated using the relevant information formalism. At last, multi-agent relevant information is set in relation to the self-organization of agent collectives and the implications of agent coordination are studied in the context of shared control.

Chapter 6 discusses the problems with currently available approaches to quantify redundant information between random variables with respect to another variable. A new measure of bivariate redundant information is developed and it shown that it resolves most of the earlier discussed problems and can be used in the decomposition of mutual information. Furthermore, an extension of the measure to the multivariate setting is proposed.

1.3 CONTRIBUTION

Parts of this thesis have been published in the following articles

- Harder, M., Salge, C. and Polani, D. (2013). Bivariate measure of redundant information. *Physical Review E, 87*(1), 012130.

- Harder, M. and Polani, D. (2012). Self-organizing particle systems. *Advances in Complex Systems*, pp. 1250089.

- Harder, M., Polani, D. and Nehaniv, C. L. (2010). Two agents acting as one. In Fellermann, H., Dörr, M., Hanczyc, M., Ladegaard, L. L. and Maurer, S. et al., editors, *Artificial Life XII: The 12th International Conference on the Synthesis and Simulation of Living Systems*, pages 599–606.

Furthermore, the following article was published during the programme without being included in this thesis

- Harder, M., Polani, D. and Nehaniv, C. (2011). Think globally, sense locally: From local information to global features. In *Artificial Life (ALIFE), 2011 IEEE Symposium on*, pages 70–77.

The contributions of this thesis are

- A comprehensive literature review on currently available methods for the estimation of multi-information in the continuous domain.

- A quantitative comparison of existing multi-information estimators in high-dimensional continuous domains.

- A framework to compute observer self-organization in spatial collective systems.

- An investigation of particle dynamics and the impact of type differentiation and interactions on the expressed amount of self-organization.

- An additional axiom for bivariate measures of redundant information.

- A novel bivariate measure of redundant information based on information geometric projections which is consistent with the partial information decomposition of mutual information (Williams and Beer, 2010) and can be easily computed using numeric optimization techniques.

- The extension of the bivariate measure of redundant information to a multivariate measure that also uses information projections and fulfils the required axioms, however

is not easy to compute and it remains open whether it is consistent with regard to the partial information decomposition of multivariate mutual information.

- The extension of the relevant information formalism (Polani et al., 2006) to multi-agent scenarios, in particular with episodic morphogenetic tasks of guided self-organization.

- A theoretical limit on the amount of collective self-organization depending on the information processing performed by individual agents of the collective as a generalization of information-theoretic limits to control by Touchette and Lloyd (2004).

- A formulation of efficient shared control by comparing information processing and actuator coordination with the achieved performance of a collective regarding a particular task.

BACKGROUND

» *I just wondered how things were put together.* «

CLAUDE SHANNON, Unknown

2

2.1 INFORMATION THEORY

Information Theory was initially introduced by Shannon (1948) to provide a theory to the problem on how to quantify and optimize transmission rates in noisy communication channels. Many of the mathematical ideas and tools used for this by Shannon (1948) stem from works on probability theory and thermodynamics by Kolmogorov (1946), Boltzmann (1866) and Gibbs (2010) and many others. Even though thermodynamics and statistical mechanics share the same mathematical foundations as information theory does, it took a little more than a decade until the first connections between information theory and physics were drawn by physicists as Landauer (1961,1991) and Jaynes (1957). A first step in this general direction was made by Wiener (1948) in the same year as information theory was formally introduced (Shannon, 1948). In his work Wiener already draws the connection between communication, control, robotics, social complexity and living organisms and one is tempted to say that his thoughts were ahead of his time.

Nowadays, after the rise of new scientific fields (complex systems, computational mechanics, quantum computing, systems biology, to name a few) and a digital revolution in society, information theory has grown in importance. Applications can be found in a myriad of disciplines (see (Attneave, 1959, Jade and Sarkar, 1993, Rashid et al., 2002, Topp et al., 2013 and Effenberger, 2013) for a small selection of applications to different fields), but it still serves its original purpose as a theory of communication underlying much of the technology that most of us are using daily to communicate with each other.

The main measure used in information theory is called *entropy* and it measures the uncertainty of the outcome from a set of possible events. Another interpretation of entropy is the average length of a symbol (in bits) when data is compressed in the best possible way. This has apparent applications for communication and compression. But in practice, any probabilistic model can be examined using information-theoretic methods, and it will be seen in Section 2.2 that this often provides new insights that would not have been accessible otherwise.

2.1.1 *Foundations of Information Theory*

Shannon (1948) introduced information theory as a mathematical theory of communication and stressed that "semantic aspects of communication are irrelevant to the engineering problem" (Shannon, 1948, p. 1). This means that the foundations of information theory do not require any semantics of the information dealt with. This is true for the "engineering problem", but even more so for the mathematical basis of information theory. However, this does not mean that information theory cannot be set in relation to semantics. Semantics

is simply not an inherent property of information in the information-theoretic sense and need to be incorporated by adding context. For example, the concept of relevant information introduced by Polani et al. (2006) provides an information-theoretic measure of 'meaningful' information in a specific context.

Information theory itself is agnostic to the context it is used in. The initial conceptual coupling with the construct of the communication channel consisting of Source, Transmitter, Channel, Receiver and Destination (Shannon, 1948) presented it very much as an engineers' tool. Thus, despite its connections to physics, early ideas about using information theory to assess biological and cognitive systems using information theory were dismissed on the basis that living organisms do not communicate with the world around them and especially not the other way around (Gibson, 1986). Here, it is important to stress that the engineers' model of a communication channel already defines the semantic context it is used in, namely sender and receivers acting with intent, even though the semantics of the information that is transmitted are ignored in this particular context.

Not only the concept of having a sender and receiver acting with intent is generally misleading, but also thinking of entities that share information as observers or agents is generally not correct. Information is a stochastic observer-independent and non-causal quantity. As it turns out later, it is however possible to incorporate observers and causality in the specific model that is used, in the same spirit as done with semantics.

2.1.2 *Random Variables & Probabilities*

Before I introduce entropy and its related measures, I will give an overview of the notation used. Random variables are denoted by italic capital letters e.g. X, Y or Z. Capital letters are also used to denote index sets and power sets, for which the letters A, B, V and R are reserved. Random variables are as usual defined via probability spaces (see (Klenke, 2008) for details). Let $(\Omega, \mathcal{F}, P)$ be a probability space, with sample space Ω, a set of events $\mathcal{F}$ and the probability measure P. An $(E, \mathcal{E})$-valued random variable on a measurable space $(E, \mathcal{E})$ is a measurable map $X : \Omega \to E$. Unless otherwise specified, random variables used in this thesis are finite. Furthermore, if no measurable space is explicitly given, a random variable X is considered to be $(\mathfrak{X}, \mathcal{X})$-valued where $\mathfrak{X}$ is a finite set of atomic events and $\mathcal{X} = \mathcal{P}(\mathfrak{X})$ is the σ-algebra generated by the powerset of $\mathfrak{X}$. Real valued random variables map into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra of $\mathbb{R}$.

Now the probability measure $P_X := P \circ X^{-1}$ is called the *distribution* of X. The cumulative distribution function F_X of a real valued random variable X is defined as $F_X(x) := P(X \leq x)$. Respectively its probability density function is denoted f_X and

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{2.1}$$

If $\mathfrak{X}$ is a discrete set and X an $(\mathfrak{X}, \mathcal{X})$-valued random variable, then $p_X : \mathfrak{X} \to [0,1]$ denotes the probability mass function associated with the distribution P_X and $p_X(x) := P_X(\{x\}) = P(X^{-1}(\{x\}))$ for all $x \in \mathfrak{X}$. I will use $p(x)$ as a shorthand for $p_X(x)$ if it is clear from the context which distribution and probability mass function is meant. The space of all possible probability distributions P_X on $(\mathfrak{X}, \mathcal{X})$ is denoted $\Delta(X)$. In the finite case, $\Delta(X)$ is a $|\mathcal{X}| - 1$ dimensional simplex.

A joint distribution of several random variables, e.g. X, Y, Z is denoted as $P_{X,Y,Z}$. Respectively a conditional distribution is denoted as $P_{X|Y,Z}$. In the same way joint or conditional probability mass functions as well as joint or conditional probability density functions will be denoted. Again, if it is clear form the context I will omit the indices denoting the random variables and simply use $p(x|y)$ instead of $p_{X|Y}(x|y)$ to denote the value of a conditional probability mass function.

As the random variables used in this thesis are mainly finite and real valued there is no need to be concerned about transition kernels and other concepts from the measure theoretic foundations of probability theory. The underlying probability space $(\Omega, \mathcal{F}, P)$ is implicitly assumed to be expressive enough to capture all the desired random variables and their distributions and thus is not further specified.

2.1.3 *Bits & Entropy*

Suppose X is a random variable, then its entropy is defined as

$$H(X) := -\sum_{x \in \mathfrak{X}} p(x) \log p(x), \tag{2.2}$$

where the logarithm is to the base 2 and $0 \log 0 = 0$ by convention. Entropy is measured in bits and quantifies the average amount of uncertainty about the outcome of X. An analogous explanation is that it measures the amount of symbols from a binary alphabet that are needed on average to transmit an outcome of X with the best possible compression. Entropy is non-negative and becomes maximal for uniformly distributed random variables $H(X) = \log |\mathfrak{X}|$. On the other hand for any distribution where there is an x, such that $p(x) = 1$, the entropy vanishes.

The conditional entropy of X given Y, is defined as the difference between joint entropy and the entropy of the conditioned variable:

$$\begin{aligned} H(X|Y) &:= H(X,Y) - H(Y) && (2.3)\\ &= -\sum_{x,y} p(x,y) \log p(x|y). && (2.4)\\ &= -\sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y). && (2.5) \end{aligned}$$

This measure is also non-negative and can be interpreted as the remaining average uncertainty of the whole system of two variables if the result of one is known. As entropy is only

additive if X and Y are independent, that means $p(x,y) = p(x)p(y)$ for all $x \in \mathfrak{X}, y \in \mathfrak{Y}$, it follows that $H(X|Y) = H(X)$ if and only if X and Y are independent.

2.1.4 *Mutual-Information*

There are several definitions of information (Floridi, 2010 and Dretske, 1981), some even include the meaning (semantics) of information. I will not be concerned with such a definition here, as explained Section 2.1.1. In information theory, information is defined as a reduction of uncertainty. This is a quantitative and non-semantic definition. Therefore, the mutual information between X and Y is defined as the difference between the entropy $H(X)$ and the conditional entropy $H(X|Y)$.

$$I(X;Y) := H(X) - H(X|Y) \tag{2.6}$$

$$= H(Y) - H(Y|X) \tag{2.7}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.8}$$

It is obvious from the above definition of conditional entropy (2.3) that mutual information is symmetric. Moreover, because $H(X|Y) \leq H(X)$ it is also non-negative. Mutual information is the information that X and Y share about each other. This is also what I mean when I speak of a variable containing information about another. It is quite easy to prove that $I(X;X) = H(X)$ and hence, entropy is sometimes called self-information.

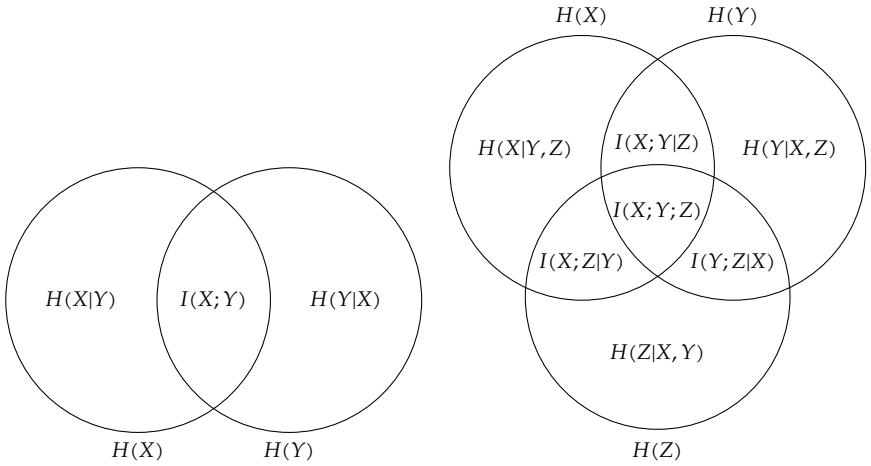
If $p(x|y)$ describes the communication over a communication channel with Y being the source and X the output, then the mutual information $I(X;Y)$ denotes the channel capacity, the amount of information that is transferred on this channel when the input has the distribution $p(y)$. The relation between entropy, conditional entropy and mutual information is illustrated using a Venn diagram Figure 2.1.

It is possible to condition mutual information on a third random variable Z, resulting in the non-negative measure of conditional mutual information:

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z) \tag{2.9}$$

$$= \sum_{z} p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}. \tag{2.10}$$

While $H(X|Y) \leq H(X)$, the respective inequality does not hold for the relation between mutual information and conditional mutual information. Thus, $I(X;Y) - I(X;Y|Z)$ may be positive, negative or zero. This means that knowing the outcome of a third variable can decrease the information shared by X and Y. The difference $I(X;Y) - I(X;Y|Z)$ is sometimes called interaction information (McGill, 1954) or co-information (Bell, 2003) and denoted by $I(X;Y;Z)$ (sometimes also the negative $I(X;Y|Z) - I(X;Y)$ is used for a definition). Some sources use the term *multi-information* for $I(X;Y;Z)$. However, I will



a) Relations between two variables X and Y.

b) Relations among three variables X, Y and Z.

FIGURE 2.1 Illustration of relations between (conditional) entropy and (conditional) mutual information.

use the term *multi-information* solely to denote *multivariate mutual information* as will be introduced below. Interaction information is obviously symmetric in all three variables and measures how 'entangled' the three variables are. If it is negative, there is a lot of 'entanglement' while in the case of positive interaction information, often one of the variables is a common cause or consequence of the other two. I will show Chapter 6 that this is not always strictly true and there are cases in between where this naive interpretation does not work.

2.1.5 *Kullback-Leibler Divergence*

There is another important measure in information theory, called Kullback-Leibler (KL) divergence. It differs from the previous three as it does not measure a quantity from a single distribution of one or more random variables, but it measures the divergence between two distributions defined on the same measurable space. Let P_X and Q_X be two distributions of $(\mathfrak{X}, \mathcal{X})$-valued random variables, then the KL-divergence between P_X and Q_X is defined as

$$D_{\mathrm{KL}}\left(P_X \,\|\, Q_X\right) := \sum_{x \in \mathfrak{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)}. \tag{2.11}$$

Again, by convention $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. The KL-divergence is similar to a distance, though not being a proper metric as it is not symmetric. It measures the inefficiency when using a code that assumes a different underlying distribution: Suppose some data occurs with distribution P_X, but the code for its description is constructed assuming the distribution Q_X, then the average symbol length of the data for which the true underlying distribution is given by P_X will be $H(P_X) + D_{\mathrm{KL}}\left(P_X \,\|\, Q_X\right)$ (Cover and Thomas, 2006) (here $H(P_X)$ is written instead of $H(X)$ to address the existence of several $(\mathfrak{X}, \mathcal{X})$-valued distributions).

Interestingly, the mutual information between X and Y can now be defined as the KL-divergence between the joint distribution $P_{X,Y}$ and the distribution $Q_{X,Y}$, which probability mass function is defined via the marginals of X and Y, i.e. $q_X(x,y) := p_X(x)p_Y(y)$. Therefore mutual information measures the inefficiency of assuming X and Y being independent.

2.1.6 *Important Properties*

There are many equalities and inequalities in information theory and specifically a whole theory regarding information-theoretic inequalities has recently been developed (Yeung, 2008). Here I will only present a few, specifically those that will be needed throughout this thesis. First of all, entropy obeys a chain rule equality. For random variables $X_1, ..., X_n$ the following is true:

$$H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1), \tag{2.12}$$

where the term in the sum reduces to $H(X_1)$ for $i = 1$. A similar equality holds for mutual information between the X_i and another random variable Y:

$$I(X_1, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, ..., X_1). \tag{2.13}$$

Now, let $X \to Y \to Z$ be a Markov chain of random variables, then

$$I(X;Z) \leq \min\{I(X;Y), I(Y;Z)\} \tag{2.14}$$

which is called the *data-processing inequality*. The idea is that each processing step limits the information that input X and output Z share if there is no other (hidden) channel between X and Z. The Markov condition, i.e. $X \to Y \to Z$ being a Markov-chain, states this now by guaranteeing that the only information channel between X and Z goes through Y.

Another important inequality is the convexity of the KL-divergence: Let (P_1, Q_1) and (P_2, Q_2) be pairs of distributions of $(\mathfrak{X}, \mathcal{X})$-valued random variables, then for $\lambda \in [0,1]$

$$\begin{aligned} D_{\mathrm{KL}}\left(\lambda P_1 + (1-\lambda)P_2 \,\|\, \lambda Q_1 + (1-\lambda)Q_2\right) \leq \lambda D_{\mathrm{KL}}\left(P_1 \,\|\, Q_1\right) \\ + (1-\lambda) D_{\mathrm{KL}}\left(P_2 \,\|\, Q_2\right). \end{aligned} \tag{2.15}$$

The proofs for these properties as well as a thorough introduction to information theory can be found in (Cover and Thomas, 2006).

2.1.7 *Multi-Information*

Multi-information is one of several possibilities (James et al., 2011) to extend mutual information to more than two variables. For random variables $X_1, ..., X_n$, multi-information can be defined as a difference of entropies or via a KL-divergence, similar to mutual information:

$$I(X_1,...,X_n) := \left(\sum_{i=1}^{n} H(X_i)\right) - H(X_1,...,X_n)$$
$$= \sum_{x_1,...,x_n} p(x_1,...,x_n) \log \frac{p(x_1,...,x_n)}{p(x_1)\cdots p(x_n)}.$$

This quantity measures the correlation between more than two variables. It is also non-negative, and will later play an important role in the assessment of self-organization.

2.1.8 *Differential Entropy*

Most of the concepts for discrete random variables can be transferred to continuous random variables. Let X be a continuous random variable where $\mathcal{X} \simeq \mathbb{R}^n$ and f its probability density function (pdf), then *differential entropy* is defined analogously as

$$H(X) := -\int_{\text{supp}(X)} f(x) \log f(x) dx, \tag{2.16}$$

where $\text{supp}(X)$ denotes the support of X, that is the closure of all $x \in \mathcal{X}$, such that $f(x) > 0$. It is important to note here, that this integral might not be defined for some pdfs and moreover differential entropy can be negative. Hence, it is problematic to interpret differential entropy in the same way as entropy. Mutual and multi-information can now be defined in the same fashion as they are defined for discrete distributions.

$$I(X;Y) := \int_{\text{supp}(X)} \int_{\text{supp}(Y)} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \tag{2.17}$$

and

$$I(X_1,...,X_n) := \int_{\text{supp}(X_1)} \cdots \int_{\text{supp}(X_n)} f(x_1,...,x_n) \log \frac{f(x_1,...,x_n)}{f(x_1)\cdots f(x_n)} dx_1 \cdots dx_n. \tag{2.18}$$

While differential entropy can become negative, differential multi-information is the limit of quantized multi-information terms and therefore non-negative. Again, a good introduction to differential entropy can be found in (Cover and Thomas, 2006), however I will not need much more than these definitions here. The corresponding estimators that will be introduced in Section 3.6.

2.1.9 *Causal Bayesian Networks*

A Bayesian network is a directed acyclic graph $G = (V, E)$ where the vertices $v \in V$ are indices of random variables X_v and the distribution of the joint random variable $X = (X_v)_{v \in V}$ can be written as a product of conditional distributions, conditioned on the parent variables of a vertex. That is

$$p(x) = \prod_{v \in V} p(x_v | x_{\text{parents}(v)}), \tag{2.19}$$

where $p(x_v|x_{\text{parents}(v)}) = p(x_v)$ by convention if the set $\text{parents}(v)$ is empty. Conversely, given a joint distribution of random variables $X = (X_1, ..., X_n)$ and a directed acyclic graph $G = (V, E)$ with a bijection between V and X, then (X, G) is a Bayesian network if all the nodes fulfil the Markov property, which means that it is independent of all its non-descendants if it is conditioned on all parents.

In a Bayesian network the directions of the arrows (the edges) are not given by the graph. For example in a Markov chain, which is also a special Bayesian network, the direction of all arrows can be reversed while the distribution stays the same, resulting in another Bayesian network for the same distribution. This means that a Bayesian network is not uniquely determined by its underlying distribution. In a causal Bayesian network (CBN) it is therefore required that all arrows are causal (Pearl, 2000), that means for an intervention at a specific random variable, the consequences of this intervention are only to be seen in the descendent within the CBN. With this restriction, it is possible to get a better interpretation of results that incorporate such a Bayesian network.

2.1.10 *The Perception-Action Loop*

The perception-action loop is a model that allows to study an embodied agent interacting with its environment using information-theoretic measures. It is an infinite CBN with random variables modelling the agent's sensors S_t, the agent's actuators A_t and the state of the world W_t. The index t denotes the time step and assumes a discrete or quantized model of time. The perception-action loop is a stochastic model of an agent interacting with its environment unrolled over time. The CBN of the perception-action loop is illustrated in Figure 2.2. The dynamics of the world $p(w_{t+1}|w_t)$ are determined by the policy of the agent $p(a_t|s_t)$, its sensor $p(s_t|w_t)$ and the world dynamics reacting to the agents action $p(w_{t+1}|a_t, w_t)$:

$$p(w_{t+1}|w_t) = \sum_{a_t} p(w_{t+1}|a_t, w_t) \sum_{s_t} p(a_t|s_t) p(s_t|w_t). \tag{2.20}$$

The model can be extended to systems with multiple agents and agents with memory. I will introduce details of these extensions in Chapter 5. This particular model was pioneered by Klyubin et al. (2004), Capdepuy et al. (2007a), Anthony et al. (2008), Salge and Polani (2009) and van Dijk et al. (2010), even though the insight that sensors are information processing mechanisms is not new and indeed has been in the focus of research for quite a while and earlier results in this area (Attneave, 1954, Barlow, 1959, Shepard, 1984 and Ashby, 1956) were a motivation to formally express embodied agents informationally.

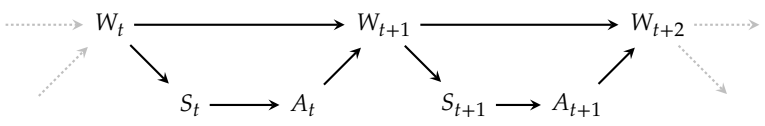


FIGURE 2.2 Illustration of the CBN of the perception-action loop of a memoryless agent.

2.2 RELATED WORK

I will now give an overview of related work drawing connections between information theory and the cognition and organization of living organisms. Towards the end of this section, a particular emphasis on self-organization and biological morphogenesis is taken, concluding in a summary of some result of recent research on information processing in actual biological embryogenesis.

2.2.1 *Embodied Cognition & Information*

The *law of requisite variety* (Ashby, 1956) states that a controller can only reduce the entropy of a controlled variable by at most the entropy of the controller. The formalization and extension of this law to modern concepts of control theory and information theory was undertaken by Touchette and Lloyd (2000,2004) and Shalizi (2001). There is a more accessible way of formulating the law of requisite variety: to reduce entropy of a controlled variable, the controller first has to acquire this information from the controlled variable. Ashby (1956) proposed this law in the 1950s and had cybernetic machines as an application in mind, but every living organism processes information from sensors to actuators, even if the boundaries between them are not always so clear. Polani (2009) took this idea and developed it even further by postulating that information might be a 'currency of life', meaning that information processing might not only be a necessary condition for life to emerge, but an existential part of defining it. In this light the perception-action loop introduced above and variations thereof become key concepts to study information processing in living systems. Moreover it becomes clear that sensors of living organisms play an important role in the study of organisms as information processing systems.

2.2.1.1 *Information Trade-offs and Relevance*

Literature on the capabilities of sensors of living organisms, as Polani (2009) noted, show that perception in living organisms often takes place in the proximity of physical limits: Human ears can operate closely at the channel capacity prescribed by thermodynamics (Denk and Webb, 1989), eyes can detect small photon clusters and for some species even single photons (Baylor et al., 1979 and Hecht et al., 1942) and it has been estimated that the photo receptors of a human eye process information in the magnitude of 10^6 bit/s (Atick and Redlich, 1992). Hence, information seems to play a very important role for living systems.

On the other hand, Atick and Redlich (1992) note that the visual pathways in the cortex only process $\sim$ 40 bit/s, which shows that a lot of the raw information is processed and compressed on the way before reaching the cortex. Barlow (2001) calls this redundancy reduction, and hypothesizes that this happens because Information processing and transmission costs energy (Laughlin et al., 1998) and energy is a limited resource for an organism.

But there can be more to it than just a redundancy reduction, as Polani (2009) emphasizes. The visual cortex is possibly not just processing compressed data coming from photo receptors of the eye but it seems that also a filtering for relevance exists (Lee and Mumford, 2003 and de Ladurantaye et al., 2012). A simple example for this is given in (de Ladurantaye et al., 2012) by noting that eye-movements already filter for relevant information "thus, the eyes strategically move to pick up relevant information for goal-directed action, and they are tightly bound to this task." (de Ladurantaye et al., 2012, p. 151). It is conjectured by Polani (2009), that the trade-off between energy and information processing makes organisms that can extract relevant parts from their sensor information perform better on an evolutionary scale. Wasting metabolic energy for information processing of non-relevant information will be disfavoured. At the same time leaving relevant information out to save metabolic energy results in the reaching of only a suboptimal policy. The implications of informational drives for sensor evolution have been discussed in (van Dijk and Polani, 2012). The picture here is simplified as adaptability may also favour a lower amount of information processing with the advantage of being less specialized but requiring less energy. Recent investigations show that metabolic constraints might further implications and for example increase the robustness of distributed learning (Balduzzi et al., 2013).

The formalization of a trade-off between information processing capabilities and performance of an agent with respect to some utility has been as far as I am aware introduced by Polani et al. (2006). The minimal amount of information needed to achieve an optimal policy with respect to some utility is called *relevant information*. Relevant information can also be defined for suboptimal policies leading to a trade-off curve which is illustrated in Figure 2.3. Relevant information is not to be confused with the similar concepts of Optimal Causal Inference (Still et al., 2007) and the information-theoretic approach to interactive learning by Still (2009). Relevant information connects to a utility that is not necessarily informationally motivated, whereas the latter concepts are concerned with trade-offs between the information that action and internal state share with the future, and the information that needs to be obtained for an internal model of the world.

Returning to the evolutionary view as taken above, all organisms operate in the achievable area and evolutionary drives work towards the limit of this area. Better performance as well as less information processing are favoured by evolution. Polani (2009) concludes this thought with the hypothesis that small variations at the limit of the achievable region enable the exploration of evolutionary advantages and thus gradually increase the organisms' performance, and therefore a drive to maximize information processing might be an evolutionary advantage. In the sense, that increasing information processing allows to explore better performing policies. If the shape of the trade-off curve is as in Figure 2.3, this will certainly be true for policies with only very small mutual information between sensor and actuator. Increasing information processing capabilities, allows an agent to reach much higher levels of performance, given that the agent is informationally parsimonious. For agents already having close to optimal performance simply increasing their information

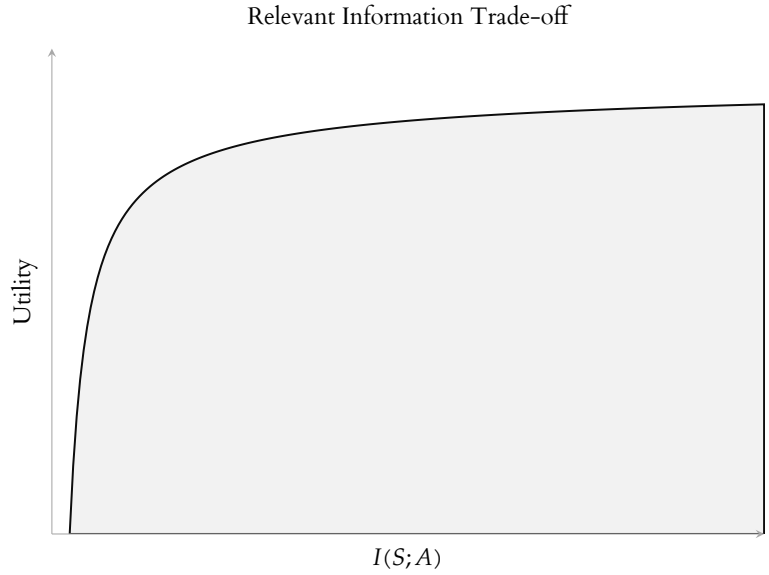


FIGURE 2.3 Illustration of the relevant information trade-off curve. To reach a higher utility more information needs to be processed. The top right corner of the curve marks the relevant information, the amount of information that is needed to follow a policy that is optimal in utility. The shape of the curve in this illustration is typical for relevant information curves of reinforcement learning scenarios: Much of the information needed by the agent only contributes to a small amount of utility at the top of the curve.

processing capabilities only gives way for small performance increases and thus in these areas a simple maximization of information processing is possibly not favoured by evolution. In general, agents being informationally parsimonious, meaning they only process relevant information, have an inherent evolutionary advantage, if the processing of information is connected to some metabolic cost.

Information maximization is a pattern that appears repeatedly in the context of cognition. It was used to optimize neural networks by maximizing the mutual information between input and output of nodes in a multi-layer neural network (Linsker, 1988) and is the basis of Infotaxis (Vergassola et al., 2007) where an artifical moth is aims to maximize the information gain about an odour source and thereby approaches the target in a biological plausible way.

2.2.1.2 *Empowerment*

Relevant information looks at the arrow from sensors to actuators in the CBN of the perception-action loop. Empowerment, another information-theoretic measure, looks at the path from actuator to sensor via the environment of an agent (Klyubin et al., 2007). It also uses a maximization approach as it is defined as the channel capacity between actuators and sensors. The empowerment of an agent can be calculated for every state of the world and gives insight into the agent's ability to perform actions whose consequences can be sensed by the agent itself at a later time. States with high empowerment often correlate with 'interesting' states and empowerment maximization can be used as an intrinsic drive to

achieve certain tasks. For example, empowerment maximization can be used to balance an inverted pendulum (Salge et al., 2013) or to produce automatic collision avoidance (Glackin et al., 2013) without explicit design of the desired behaviour. In collectives, empowerment maximization can even lead to the formation of cell-like shapes with a membrane (Capdepuy et al., 2007b).

2.2.1.3 *Predictive Information*

The perception-action loop is a very abstract model of cognition and interaction of biological systems with their environment. However, at the boundary between biology and physics the perception-action loop cannot be used it since there is no clear concept of embodiment in this region. In this case the complexity, structure and interactions of time series are important quantities to investigate such systems. There are many information-theoretic measures to study time series of random variables without the definition of a perception-action loop. One of these measures is predictive information. It is defined as the mutual information between past and future of a time series of random variables at a given time step (Bialek et al., 2001). Predictive information therefore poses an upper bound on what can be known about future observations given all past observations. This quantity is related to the complexity of the time-series (Bialek et al., 2001 and Shalizi, 2001) and has an inherent connection to learning. Part of learning is the extraction of predictive information from past observations to be able to anticipate future observations. The information between infinite past and infinite future can be defined as a limit of the sequence of information between finite pasts and finite futures. The convergence behaviour of this quantity allows conclusions about the complexity of the underlying model (Bialek et al., 2001).

In the framework of the perception-action loop, predictive information is usually measured between past and future sensor readings of an agent. Maximizing this quantity leads to interesting behaviours where agents explore their environment autonomously (Ay et al., 2008) or coupled agents begin to cooperate (Zahedi et al., 2009). The learning rules to maximize predictive information are in this case similar to the principle of homeokinesis (Der et al., 1999) where an agent minimizes the prediction error of its self-model. Predictive information, empowerment and relevant information are possibly related quantities, but not in an obvious way that one is the complement of the other nor necessarily an upper bound.

2.2.2 *Information Theory, Structure & Complexity*

Predictive information is related to the question of the complexity of a time series. There are many definitions of complexity and structure, Chapter 3 will give a detailed introduction into the information-theoretic concepts of complexity and in Chapter 6 I will take a look at

the informational structure of CBNs. Here, I will only give a brief overview and introduce some related concepts.

Early works connecting complex systems and information theory include (Bennett, 1990 and Tononi et al., 1994). The notion of statistical complexity, based on the entropy of a time series' causal state automaton, was introduced by Crutchfield (1990,1992) and later extended by Shalizi (2001). There is a close connection between statistical complexity and predictive information which allows to quantify self-organization in terms of prediction efficiency. A very good overview of many of these information-theoretic measures that can be used on a time series of random variables can be found in (James et al., 2011).

While statistical complexity is mainly used in the temporal dimension of a system, there are other measures that are helpful to investigate the structure of correlations between individual parts of a system. For example, interaction complexity (Kahle et al., 2009 and Ay et al., 2006a) is an information-theoretic measure that quantifies correlations between k parts of a system. It measures interaction between exactly k parts and gives an orthogonal view in comparison to multi-information by providing a decomposition of it into informational contributions of k-interactions.

Another important measure to mention at this point is transfer entropy (Schreiber, 2000), which is simply the mutual information between the current state of one time series and the next state of another, conditioned on the current state of the other time series. Transfer entropy provides an insight into how much one time series is influencing another. It has been extended by (Lizier et al., 2007 and Lizier, 2011) to measure local information transfer and storage in multi-agent systems (Wang et al., 2011), and cellular automata (Lizier et al., 2013).

2.2.2.1 *Redundancy, Synergy and Integration*

Informational quantities are usually averages and new methods are needed to identify information that is identical, for example to detect redundancies when the measures introduced in the last section are used to study interaction between several time series. Being able to detect identical information enables the decomposition of the structure of the information that several random variables share with another random variable. With the partial-information decomposition Williams and Beer (2010) provide a systematic approach to distinguish between redundant, unique or synergistic contributions to mutual information. It is based on lattice theory and provides a formalism once one defines the underlying measure of redundant information. This field gained a lot of traction over the recent years (Griffith, 2011, Griffith and Koch, 2012 and Bertschinger et al., 2012) and one of its applications is the structural decomposition of multivariate transfer entropy (Flecker et al., 2011 and Lizier et al., 2013). I will give a detailed introduction to this formalism in Chapter 6.

Related to the decomposition into redundant and synergistic information terms, is the measure of integrated information (Tononi et al., 1994 and Balduzzi and Tononi, 2008,2009).

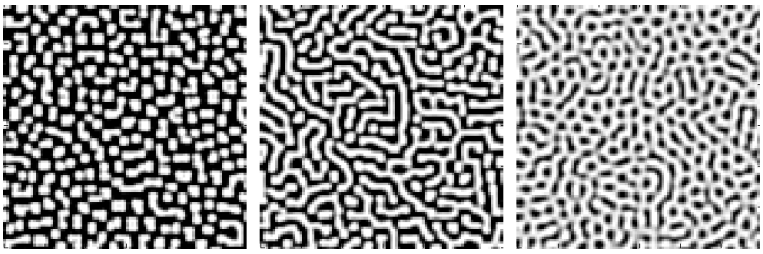


FIGURE 2.4 Examples of Turing patterns generated with diffusion-reaction systems.

It measures how much a system consisting of several parts, that evolves over time, can be decomposed into components with independent dynamics. If it is possible to partition a system into independent components, the measure vanishes and there is no integration in the system (Balduzzi and Tononi, 2008).

2.2.2.2 *Causality*

As a last remark on information-theoretic measure, I want to mention the causal information flow. While most of the measures introduced earlier only quantify correlations between random variables of a CBN, there is also a way to quantify causal influences within a CBN. Equipped with the causal structure (which is known from construction or was reconstructed using a structural learning algorithm (Pearl, 2000)), one can use the causal information flow formalism (Ay and Polani, 2008) to quantify the strength of causality in a CBN using an interventional approach similar to the *do* notation of Pearl (2000).

2.2.3 *Self-Organization*

There is a plethora of examples for self-organization and self-organizing systems which have attracted researchers from various fields in the last century. From fluid dynamics to geophysics, pattern recognition to neural networks, population dynamics to morphogenesis, all these fields deal with phenomena of self-organization.

The foundations for a quantitative analysis of self-organization are closely related to the quantification of complexity and were laid by (Kolmogorov, 1963, Chaitin, 1969, Wolfram, 1986, Feldman and Crutchfield, 1998 and Shalizi, 2001). In Chapter 3, I will give a review of the work on quantification of self-organization and will use the approach by Polani (2008) to measure self-organization via multi-information. Here, I want to review some of the actual simulations of self-organizing systems that were performed and their relation to biology.

2.2.3.1 *Simulations of Morphogenesis*

I am mainly interested in simulations of self-organizing systems for the purpose of this theses, especially multi-agent system and simulations of morphogenesis. One of the first

work in this area deals with the famous Turing-patterns (Turing, 1952), simulating pattern formation via diffusion-reaction systems as illustrated in Figure 2.4. Key works in this area are also the reaction diffusion systems by Gierer and Meinhardt (1972). They considered specific variants of reaction-diffusion models, activator-inhibitor systems, and used them to explain the formation of organising regions and primary gradients (Gierer and Meinhardt, 1972 and Meinhardt, 1982,2006). A general introduction to pattern formation in an artificial life context can be found in (Bonabeau, 1997).

A well known benchmark for morphogenesis is the French flag problem, that is, the formation of three bands of distinctly differentiated cells using a gradient of a morphogen (Wolpert, 1969, Miller, 2004 and Knabe et al., 2008). The challenge is that a cell on a local scale has to decide into which colour it needs to differentiate where the colour depends on its absolute position, which is a global property of the system.

Beyond that, more specific work has been done on morphogenesis within the artificial life community (Graner and Glazier, 1992, Theraulaz and Bonabeau, 1995, Jones, 2010 and Hogeweg, 2000). To highlight a few: Simulations of morphogenesis by Doursat (2008b,2008a) show how neural networks can be evolved to control a coordinated growth of limbs and other forms. The formation of *Physarum* transport networks has been simulated using multi-agent models (Jones, 2010). Odell et al. (1980) simulated the gastrulation stage in embryonic development using a multi-cellular model where the cell boundaries are loaded springs. Gastrulation is a key phase during embryonic development and understanding the driving mechanisms behind it is key to understand morphogenesis in general. A biologically precise simulation of gastrulation of chick embryos was done by Vasiev et al. (2010). These simulations are closer to the actual biological processes of morphogenesis, and are not only explanatory models but nowadays even serve as predictors. Where earlier models mainly explored what principles could in theory lead to certain morphogenetic processes, recent models like the work of Jakobsson et al. (2010) on vascular morphogenesis, or Bentley et al. (2005) on models of diatom valve morphogenesis, were built with the aim of being used as predictive models.

Robotics and especially nano-robotics are concerned with the problem of morphogenesis in the design of self assembling systems. Here agent based models are generally used and there are several constructive approaches to create languages that specify formation processes. Christensen et al. (2008) deployed a system to the SWARM-BOT platform (Mondada et al., 2004) that allows a collective of robots to assemble in specific shapes by passing rules between connected robots; Rosa et al. (2008) presented a language based on predicates that can be used to specify local interactions between modules of a modular robot and hence its shape. Furthermore, a group at Harvard specifically deals with the problem of self-assembly from a theoretical (Werfel and Nagpal, 2006 and Cheng, 2005) and practical (Rubenstein et al., 2012 and Yu and Nagpal, 2011) point of view, providing links between the simulations and the actual implementations.

2.2.3.2 *Biological Morphogenesis & Information*

Early studies of biological morphogenesis have been mentioned in the introduction. A well written introduction to the general field of biological development is provided by Wolpert et al. (2002). Research in this area has been quite active in the last 50 years (Townes and Holtfreter, 1955, Sinnott, 1960, Wolpert, 1969, Summerbell et al., 1973, Tickle et al., 1975, Thom, 1989, Bard, 1990, Gumbiner et al., 1996 and Davies, 2005), including research of overlapping fields like genetic regulatory networks that control the process of cell differentiation (Wolpert et al., 2002). As remarked earlier, I will not be as concerned in this thesis with the distinction between morphogenesis, cell differentation, regulation timing or patterning, as some of the biological literature does. These mechanisms are not only related, but often work together in the process of biological development. Even more so, from an abstract point of view I propose that it is sometimes impossible to distinguish between them.

As in several other areas within biology, the fruit fly *Drosophila melanogaster* is a reference organism on which much of the research of morphogenetic processes has been done (together with the African clawed frog *Xenopus laevis*, and the embryos of zebrafish, chicken and mice) (Wolpert et al., 2002). Much of the research is about understanding gene regulatory networks in connection with morphogenesis. In the *Drosophila* embryo, gene expression levels provide a 'blueprint' of body axis and segmentation. While the underlying regulatory mechanisms and maternal factors for the expression levels have already made it into the textbooks (Wolpert et al., 2002), a quantitative analysis of the regulatory mechanisms is quite recent.

Noise in gene expression has been studied in (Ozbudak et al., 2002 and Blake et al., 2003) and Dubuis et al. (2011) even use information-theoretic methods to quantify the amount of positional information that is encoded in the gene expression levels of *Drosophila* embryos. They investigated whether the information provided by the gene expression levels is enough to explain the pattern along the anterior-posterior axis of the embryo. Not only were they able to show that the information provided by the gene expressions level suffices to determine the position of a cell only with a 1% error along the anterior-posterior body axis (Dubuis et al., 2011), it was moreover possible to see that the positional uncertainty is constant over the whole span along the anterior-posterior axis. This result is quite intriguing as it is what was predicted to be the positional uncertainty for a regulatory network that is informationally optimal (Tkačik et al., 2008,2009) and therefore seems to suggest that a selection for optimal positional information processing has happened in the development of *Drosophila* embryos. Hence the result provides yet another example for biological organisms operating at the physical limits of information processing.

In another article Gregor et al. (2007) show that in *Drosophila* embryos the gene expression of the *Hunchback* gene, which is a read out of the *Bicoid* gene, has a precision close to the physical limit with the given noise constraints induced by random arrival of individual

molecules. This is not only interesting because these are additional examples for biology operating at physical limits, but also because it requires a change in thinking about the early stage of morphogenesis and the layout of body plans. It was thought (Houchmandzadeh et al., 2002 and Von Dassow et al., 2000) that understanding these processes would be much about understanding how noisy mechanisms in connection with input data of high entropy lead to precise body layouts (Gregor et al., 2007). However, it seems that the input data has surprisingly low entropy, close to the physical limit, and the actual question is, how is such a precision achieved in the first place, namely in the very first stage of embryonic development (Gregor et al., 2007).

QUANTIFYING SELF-ORGANIZATION

3

» *Living organisms are metastable Maxwell demons whose stable state is to be dead.* «
NORBERT WIENER, Cybernetics

3.1 INTRODUCTION

Although it seems quite easy for humans, from a visual standpoint, to point out whether we consider a system as self-organizing ('I know it when I see it'), I am aware of surprisingly few quantitative characterizations of self-organization that could be applied to the examples of self-organization as introduced in Section 2.2.3.1. Obviously, there is a connection between the concepts of organization, complexity, structure and patterns, even though the terms cannot be used interchangeably. There is a large body of literature on pattern recognition (see (Bishop et al., 2006) for an overview) and pattern formation (Harrison, 1994 and Bonabeau, 1997), but mostly these accounts fail to give a concise theory of what a pattern inherently is or how to detect patterns, and thus do not help in the process of detecting the emergence of patterns. Turning to treatises on biological organization does not shed any more light into the darkness, as they mainly describe specific models of self-organizing systems (Lwoff, 1962, Quastler, 1964, MacMahon et al., 1978, Meinhardt, 1982 and Weng et al., 1999) and, as Shalizi (2001) noted, sketch only what a theory of organization ought to do, but not what it should be. In this context, the work of Lwoff (1962) on 'biological order' stands out as it is one of the few earlier works already drawing a systematic connection between biology, order and entropy as well as information.

Possibly one of the earliest most explicit definitions for organization is given by Wolfram (1986), who defines organization as the reduction of thermodynamic entropy. There are several points of critique speaking against the use of entropy reduction as a measure of organization. The most fundamental argument speaking against entropy (either thermodynamic or information-theoretic) as a measure of organization comes from Bennet, who argues "the human body is intermediate in entropy between a crystal and a gas" (Bennett, 1993) and by such a process transforming a human into a crystal would be considering as organizing.

3.2 SELF-ORGANIZATION & COMPLEXITY

It seems that the intuitive definition that organization is an increase of complexity over time has many proponents (Dalenoort, 1989, Bennett, 1990, Shalizi, 2001 and Polani, 2008) and can be useful, as long as there is a good definition of complexity available. But this shifts the problem only infinitesemally forward. The question posed now is: what is a good definition of complexity? There is the well known picture of the complexity curve (Crutchfield and Young, 1989), where disorder or entropy is on the x-axis and complexity on the y-axis and complexity is maximized somewhere between periodic sequences like 'AAAAA...' and randomness on the other extreme of the x-axis (see Figure 3.1 for an illustration).

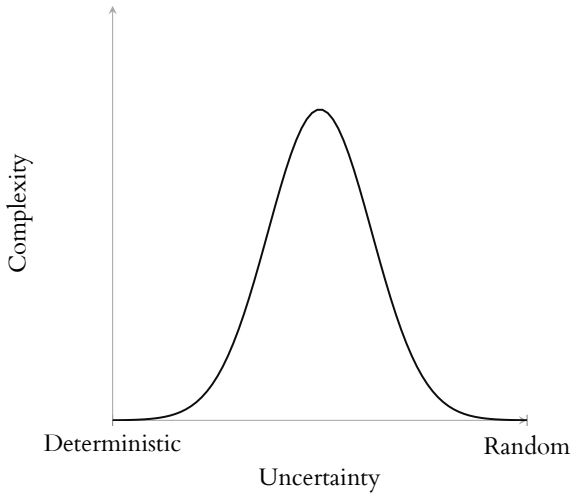


FIGURE 3.1 The "one-humped complexity curve" as introduced by Crutchfield and Young (1989), illustrating the idea that systems of high complexity lie somewhere between simple deterministic and completely random systems.

Bennett states that subjective organization is not additive and gives the example of a bacterial culture in a nutrient solution (Bennett, 1990). He proposes that growth of the bacteria does not show any more organization than what is already available in the seed bacterium. One bacterium contains more organization than no bacterium, but two identical sibling bacteria contain only slightly more than one of them (Bennett, 1993), an idea put forward already by Lloyd and Pagels (1988). Ladyman et al. remark, that these ideas basically capture the idea that "complexity [and thus organization] has something to do with how difficult it is to produce something" (Ladyman et al., 2011, p. 18).

I will now give an overview of some of the available complexity measures and discuss why most of them seem unfit in general or for the purposes of this thesis. This is by no means a complete overview of all complexity measures. A query on 'definition of complexity measure' on Google Scholar shows that such a project would possibly fill a whole encyclopaedia. A thorough review with an extensive discussion on what a complex system is can be found in (Ladyman et al., 2013), where in particular a list of several properties associated with complex systems is presented. These properties are, in short: Nonlinearity, feedback, spontaneous order, robustness, emergence, hierarchical organization and numerosity (as in many interacting entities). Without going into the detail here, this list should not be seen as a definition of a complex system, but as properties related to, implied by or necessary for complexity.

3.2.1 *Logical & Thermodynamic Depth*

Bennett, who sees self-organization as the spontaneous occurrence of a complexity increase, hence introduces a complexity measure called *logical depth* (Bennett, 1990). Similar to Kolmogorov-Chaitin complexity (Kolmogorov, 1963 and Chaitin, 1969) universal Turing

machines are used in the definition. The Kolmogorov-Chaitin complexity of a sequence of symbols from a finite alphabet is measured by the length of the minimal program on a universal Turing machine that outputs the given sequence. However, looking at the length of a minimal program makes randomly generated data maximally complex.

Thus, instead of considering the length of a minimal program, logical depth quantifies how long a program that is compressible by s bit needs to output a sequence. In this context compressible by s bit means that the minimal program generating the considered sequence is s bit shorter. For example, for an incompressible string of randomly generated symbols, the shortest program simply copies the string to the output, which is the fastest way to produce the data. On the other hand if it is possible to compress the string and the runtime of the program is longer (compared with the uncompressed version that simply copies the string to the output), the string is considered to be more complex.

This measure directly operates on sequences and not on statistical ensembles. However, the definition uses Kolmogorov-Chaitin complexity which is non-computable (Li and Vitanyi, 1993) and is therefore not practical to use with empirical data, as it is the case for any other measure based on Universal Turing Machines like for example effective complexity (Gell-Mann and Lloyd, 2004).

On the other side there are complexity definitions like thermodynamic depth (Lloyd and Pagels, 1988) which are computable, however turn out to be merely a measure of complexity but randomness (Crutchfield and Shalizi, 1999 and Ladyman et al., 2013).

3.2.2 *Statistical Complexity*

One point that was brought forward by Crutchfield (1992) and Shalizi (2001), is that most complexity measures cannot distinguish between qualitatively different kinds of organization, they are called over-universal. The lack of a well-behaving (in terms of randomness) and non over-universal measure was the motivation for Crutchfield and Young (1989) and later Shalizi (2001) to develop new ways of quantifying complexity which lead to the development of an elegant framework to study the structure of a process, which can also be used for the classification of processes. The complexity measure within this framework is called statistical complexity and will be described in more detail Section 3.3.

3.2.3 *Information as Organization*

The idea that there is a connection between mutual information and organization has been around for a while. Bennett (1993) already mentions that "subjectively organized objects generally have the property that their parts are correlated". He considers *long range mutual information*, i.e. the correlation between remote parts of a system, as a measure of organization, but dismisses it for several reasons. First of all, he states that some not very organized objects like igneous rock or other polycrystalline solids have long range

mutual information through crystal defects or grain boundaries. The second argument is, that mutual information does not obey the law of slow growth. He gives the example of an ordinary glass that is worked with a hammer contains quickly more remote mutual information than a genome.

With respect to the first point, I am adopting the position that the process in which for example grain boundaries are formed should be considered as organization, as long as there is an actual increase of information between remote parts of the system, they just might not organize as much as for example a growing plant does.

The misconception of the latter point however is, to assume that the shattering of the glass introduces a lot of correlation and not just mainly an increase of entropy in all parts of the system. Small variations in the points of impact, noise in the dynamics, will lead to completely different configurations so that there will be almost no correlation between the position and shape of the shards. If there is no variation in the system, i.e. a deterministic system with a single starting state it does not make sense to employ information-theoretic methods in the first place as there is no phase space with variations to study. I would propose the view that not the complexity measure has to obey the law of slow growth, but that given the locality of interactions in physics complex systems usually obey the law of slow growth.

3.2.4 *What is Needed for Self-Organization besides Complexity?*

Before I continue to present how self-organization can be defined formally in the context of a complexity measure, I want to discuss what 'self' denotes within the term 'self-organization'. The idea is, that there is a difference between the observation of organization and an actually self-organizing system in the sense of "self-generated complexity" (Grassberger, 1986). First consider what self-organization is not: Most importantly, there should not be an external or central force that controls the process in a 'top-down' fashion. Ladyman et al. (2013) assert that lack of central control is an inherent property of a complex system, however this does not explain how to distinguish a complex system from a system that seems complex but is actually causally dependent (Pearl, 2000 and Ay and Polani, 2008) on another self-organizing system, but is not complex itself. The dependent system should not be considered as *self*-organizing. This means that a self-organizing system needs to be autonomous, for which there is no obvious characterization (Bertschinger et al., 2008). External influences need to be separable in a quantitative manner, which can be investigated using the measure of causal information flow (Ay and Polani, 2008). The focus on the distinction between organization and *self*-organization was emphasized by Polani (2008), who also proposed the information-theoretic methods mentioned here for the study of this specific question. However, I will not consider these problems here, they are more apparent if the underlying model is unknown. In this thesis, I will only look at systems where the model is known or designed to exclude any unknown external influences.

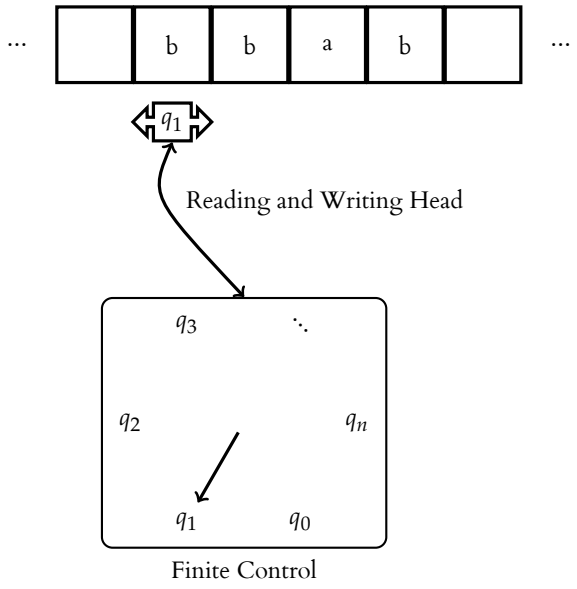


FIGURE 3.2 Illustration of a (universal) Turing machine (based on an illustration from http://texample.net by Ludger Humbert licensed under CC-BY 2.5). A Turing machine consists of a finite control (finite number of states) a reading and writing head and an infinite tape consisting of symbols of an alphabet and a set of transition rules from state and input to a new state and a tapte shift (Turing, 1936). A Universal Turing Machine is a Turing Machine that first reads in the description of a Turing Machine and then simulates this Turing Machine on arbitrary input.

This still leaves the question of how to quantify organization. As mentioned above, increase of complexity over time is a common definition for organization and I will use it as well. Its usefulness and quality still depends on the used measure of complexity. It is noteworthy that the definition also captures an increase of structure over time as organization, given that structure and hierarchies, like for example spatial correlations between parts of the whole system, are reflected in the complexity measure. I will now give two definitions of self-organization, the first one by Shalizi (2001) and the second by Polani (2008) to then discuss their respective areas of application.

3.3 STATISTICAL COMPLEXITY

Crutchfield and Young (1989) introduced statistical complexity by extending the definition of Kolmogorov-Chaitin complexity (Kolmogorov, 1963 and Chaitin, 1969). The measure of Kolmogorov-Chaitin complexity itself is defined as the size of the minimal representation (denoted M_r) of a sequence x where the representation is a program for a Universal Turing Machine (UTM, see Figure 3.2) that outputs x

$$K(x) := \|M_r(x|UTM)\| . \tag{3.1}$$

It is possible to replace the UTM by another class of languages, one situated lower in Chomsky's hierarchy of languages (Chomsky, 1956), for example Deterministic Finite Automata (DFA) or Stack Automata (SA) resulting in another complexity measure, which necessarily

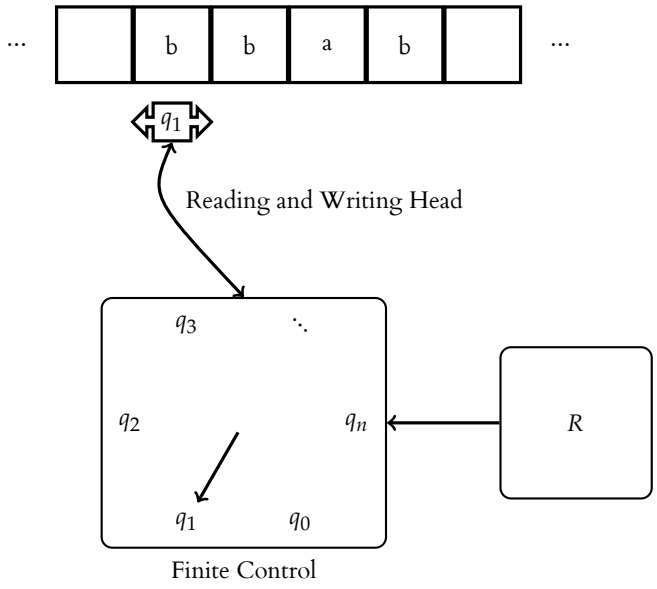


FIGURE 3.3 Illustration of a Bernoulli Turing Machine (based on an illustration from http://texample.net by Ludger Humbert licensed under CC-BY 2.5). A Bernoulli Turing Machine is a Universel Turing Machine which transition rules have access to a register containing true (thermodynamical) randomness.

measures a larger or equally large complexity, due to the more limited expressiveness of the language than UTMs. Crutchfield and Young (1989) replace the UTM by a Bernoulli Turing Machine (BTM, see Figure 3.3) which is a Universal Turing machine that has an additional instruction to sample a random register. This allows to generate periodic as well as purely random sequences with a very small program on the BTM. Hence, they define statistical complexity as

$$C_\mu(x) := \|M_r(x|BTM)\| \,. \tag{3.2}$$

This definition now switched context, from x denoting a sequence of symbols to x denoting a random process. Furthermore, it is still quite impractical, as there is no algorithm for finding a minimal BTM (or even UTM for that matter) representation of a random process. But it is possible to go down in the hierarchy of languages and equip those less expressive languages with a stochastic component. Crutchfield and Young (1989) and Shalizi (2001) do this with DFA and SA, by assigning probabilities to the transitions, in the former case leading to Stochastic Deterministic Automate SDFA (not to be confused with Non-deterministic Finite Automata, NFA). Now it is possible to reconstruct the minimal SDFA for empirical data of a conditonal stationary stochastic process. A *stochastic process* is defined with respect to the underlying probability space $(\Omega, \mathcal{F}, P)$ as follows: Let $I \subset \mathbb{R}$ an arbitrary index set, then a stochastic process is a family of random variables $X = (X_t, t \in I)$ (on $(\Omega, \mathcal{F}, P)$) with values in a measurable space $(E, \mathcal{E})$. Here I will assume that $I = \mathbb{Z}$. Now, let $\overleftrightarrow{X} = ..., X_{-2}, X_{-1}, X_0, X_1, X_2, ...$ denote a conditionally stationary stochastic process of countable random variables X_t. Conditional stationarity of $\overleftrightarrow{X}$ means

that the empirical transition probabilities are time translation invariant. Furthermore, $\overleftarrow{X}_t$ denotes the semi-infinite past at t $\ldots, X_{t-2}, X_{t-1}$ and $\overrightarrow{X}_t$ the semi-infinite future at t given by $X_t, X_{t+1}, \ldots$. Now the states of the minimal SDFA representing $\overleftrightarrow{X}$ are called *causal states* and are denoted by $\mathcal{C}$. Moreover, there is also a map from all historical realizations $\overleftarrow{x}$ (at any time) of the process to a causal state

$$\epsilon(\overleftarrow{x}) : \overleftarrow{\mathcal{X}} \to \mathcal{C}. \tag{3.3}$$

The map ϵ and the transition matrix of the SDFA, denoted T_C, is now called the ϵ-machine of the process (Crutchfield, 1994 and Shalizi, 2001). The epsilon machine now determines how histories of the process are mapped to causal states and represents the dynamics of the causal states. This induces a random variable C on the causal states $\mathcal{C}$ which is now used to define the statistical complexity of the process $\overleftrightarrow{X}$ as

$$C_\mu(\overleftrightarrow{X}) := H(C). \tag{3.4}$$

A nice property of this measure is that, it is possible to reconstruct ϵ-machines from empirical data of random processes and the software for doing so is available under an open source license (Shalizi and Klinkner, 2004).

3.3.1 *Definition of Self-organization via Statistical Complexity*

To use statistical complexity for the definition of self-organization, a minor tweak of the measure is needed. In the current form, it measures the complexity of the whole process, but for a definition of self-organization an increase of complexity over time needs to be observed (Shalizi, 2001). Instead of using a steady-state or equilibrium distribution underlying the causal states C, an initial distribution of the causal states at the starting time $t = 0$ is taken and the distribution is evolved over time with the transition probabilities of the SDFA (Crutchfield, 1992). Now, the statistical complexity at time t can be defined as

$$C_\mu(t) := H(C_t) \tag{3.5}$$

where the probability mass function of C_t is defined by

$$p_{C_{t+1}} = p_{C_t} T_C. \tag{3.6}$$

This gives way for a formal definition of (self-)organization.

Definition 1. (Shalizi (2001)) *A system (random process) has organized between time t and time $t + T$ if and only if $C_\mu(t) < C_\mu(t + T)$.*

In the original definition the word self-organization is used and the system is required to be dynamically autonomous. Many of the investigated systems can be easily identified as autonomous, some even are simply so because they were modeled in such a way. In all

other cases it is still possible to speak of organization and I will refer the reader to look into measures of causal information flow (Ay and Polani, 2008) and information-theoretic autonomy (Bertschinger et al., 2008) with regard to the question of determining how much of the organization actually comes from within.

Measuring self-organization via statistical complexity has the limitation and advantage that it assumes no structure on the space underlying the time-series. It is an advantage because the measure is very versatile and does not need to be changed for different spatial or compositional structures. On the other hand it is a limitation, because the structure of the environment will be implicitly encoded in the states of the ϵ-machine, which makes it less accessible for further analysis (Shalizi, 2001). There is an extension of ϵ-machines called spatial ϵ-machines which have a structural assumption about the space they are defined on. This gives rise to a spatial measure of statistical complexity (Shalizi et al., 2004). The spatial version of statistical complexity requires a definition of light cones, which are areas in space and time that are affected from a given point in space and time. Statistical complecity was measured for cellular automata (Shalizi et al., 2004) and random Markov fields (Shalizi and Shalizi, 2003) using the spatial version of statistical complexity. However, this measure is limited to discrete variables in practice. For continuous spatial systems I will present another measure of self-organization in Section 3.4.

3.3.2 *Predictive Information and Emergence*

There is a distinction between emergence and self-organization, emphasized by Shalizi (2001) and Polani (2008), which has been often overlooked. Depending on the definition of self-organization there are different definitions of emergence (Shalizi, 2001 and Polani, 2004).

To understand the definition of an emergent process as introduced by Crutchfield (1994), I need to introduce the measure of predictive information first. *Predictive information* is the mutual information shared between the semi-infinite past of a process with the semi-infinite future of a process, i.e. in the notation as above

$$\mathbf{E}(\overleftrightarrow{X}) := I(\overleftarrow{X}; \overrightarrow{X}). \tag{3.7}$$

This quantity, sometimes also called *excess entropy*, was extensively studied by Crutchfield and Feldman (2001), who investigated its relation to entropy rates and block entropy in random processes. Predictive information was also proposed as a complexity measure by Grassberger (1986) under the name of *effective measure complexity*. It is shown in (Shalizi, 2001), that the predictive information is always less or equal than the statistical complexity of a process:

$$\mathbf{E}(\overleftrightarrow{X}) \leq C_\mu(\overleftrightarrow{X}). \tag{3.8}$$

The difference between both quantities can be seen as the overhead of storage/computation capacity that is needed to reduce the semi-infinite pasts to a single causal state which in turn defines the future of the process. Causal states are similar to a bottleneck variable in the information bottleneck method (Shalizi and Crutchfield, 2002). The information bottleneck is a general framework for constrained optimization based on mutual information quantities in CBNs (Tishby et al., 1999). In this case the bottleneck is taken between past and future, that means the bottleneck variable maximizes the information about the future while minimizing or being constraint by the information from the past. Similarly to predictive information being smaller than statistical complexity, 'squeezing' information into a bottleneck usually creates a variable with larger entropy than the mutual information between the original variables.

The prediction efficiency is now defined as the fraction of the complexity that is actually used to predict the future

$$e(\overleftrightarrow{X}) := \frac{\mathbf{E}(\overleftrightarrow{X})}{C_\mu(\overleftrightarrow{X})}. \tag{3.9}$$

Prediction efficiency therefore is the amount of "historical memory stored in the process which does 'useful work' in the form of telling [us] about the future" (Shalizi, 2001). A process is now called *emergent* if there is a functional relation to another process that results in a higher prediction efficiency.

Definition 2. (Shalizi (2001)) *A random process $\overleftrightarrow{X}'$ is called emergent if there exists a function f such that $X_t' = f(X_t)$ for all t and $e(\overleftrightarrow{X}') > e(\overleftrightarrow{X})$.*

An interesting remark is that an emergent process can be identified with a grouping of sub-machines of the ϵ-machine of the original process (Shalizi, 2001). A sub-machine is a coarsening of the causal states by grouping strongly connected components of its SDFA (that is there are transitions from each state to each other of the component), while leaving the transitions between the grouped states consistent and deterministic.

This is one of the few formal definitions of emergence and it follows the idea that emergent descriptions are those that simplify the understanding of the process by looking at a higher level, but without including any additional information about the process (i.e. solely depending on a lower level description). As the predictive information $\mathbf{E}(\overleftrightarrow{X})$ cannot be increased using a derived process via a function f, a simplification can only be reached by lowering the complexity of the process and possibly losing a smaller amount of predictability. It can already be seen here that emergence and self-organization in this definition are distinct concepts, not necessarily implying each other, even if in many complex systems both occur at the same time, i.e. self-organization often has an emergent description and therefore is listed in (Ladyman et al., 2013) among the properties that are associated with complex systems.

In literature on self-organization and emergence, the word emergence is often used in the sense of 'patterns emerging over time' and not as introduced above as 'emergent description of a process'. However, it is not hard to come up with a time dependent version of prediction efficiency $e(t)$, where an emerging pattern would be captured by an increase of the difference of the prediction efficiency of the emergent process $e'(t)$ and the underlying one $e(t)$ over time.

3.4 SELF-ORGANIZATION VIA OBSERVERS

I will now introduce a different measure of self-organization that was first advocated by Polani (2002). It is based on multi-information and has some interesting applications for spatial systems or systems that have some clearly identifiable components. Self-organization is now considered to be an increase of multi-information between observer variables $I(X_1^{(t)}, ..., X_n^{(t)})$ as introduced by Polani (2002). Observers are defined as follows:

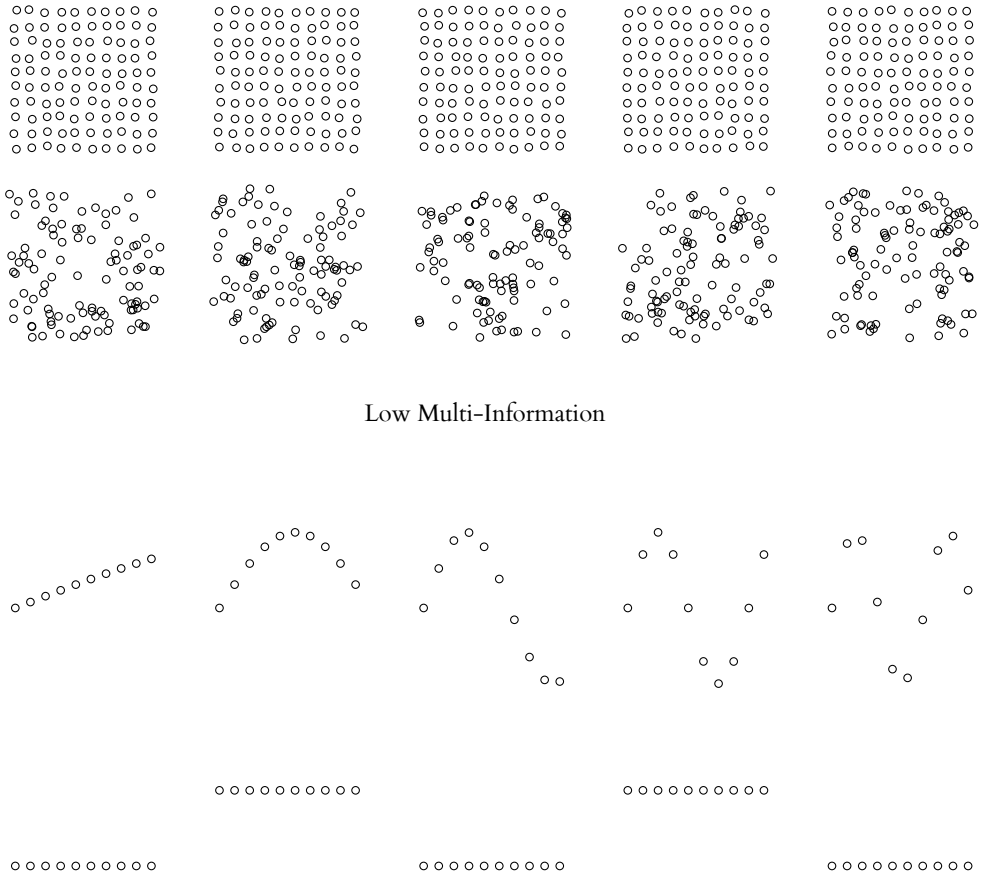
Definition 3. (Polani (2002)) *A collection of random variables $X_1^{(t)}, ..., X_n^{(t)}$ are called observers of a system $X^{(t)}$ if the system is fully determined by the collection of random variables.*

Where fully determined means that $H(X^{(t)}|X_1^{(t)}, ..., X_n^{(t)})$ vanishes, and all variables only depend on $X^{(t)}$, meaning $H(X_1^{(t)}, ..., X_n^{(t)}|X^{(t)})$ also vanishes. As long as they fulfil this condition, the observers impose a 'coordinate system' on the observed random variable and can be chosen freely. Now organization, in the sense of Polani (2002), is formally defined as

Definition 4. (Polani (2002)) *A system $X^{(t)}$ has organized between time t and $t + T$ if and only if $I(X_1^{(t+T)}, ..., X_n^{(t+T)}) > I(X_1^{(t)}, ..., X_n^{(t)})$ for some observers $X_1^{(t)}, ..., X_n^{(t)}$ of $X^{(t)}$.*

Specifying observers is an additional assumption. This is not very problematic in practice, as there are often natural choices for observer variables, as many systems are just a collection of individual random variables.

For a uniform random process, this measure never detects any self-organization, as there is no increase in correlation between observer variables. Note that there could be correlation between observers because they observe the same part of X. The other extreme case is that the entropy of the whole system vanishes, e.g. an attractor, in which case there also cannot be an increase of multi-information between the observers. So, to achieve self-organization, the system requires, besides the earlier mentioned autonomy of the process, which I consider as intuitively given by the system's isolation, without defining it in detail, some remaining degree of freedom. In the following I will adopt the labels used by Polani (2008) and call organization via observers O-organization and organization defined via statistical complexity SC-organization. Unless otherwise stated, organization, without further qualification, always denotes O-organization from here on.



Low Multi-Information

High Multi-Information

FIGURE 3.4 Illustration of systems exhibiting different amounts of multi-information. Each row depicts samples from different distributions of points in the unit square. The first two rows illustrate the extreme cases of low multi-information (between the random variables of the positions of individual points), in the top row there is no variation at all and in the second row there is no correlation between the points. The bottom rows show samples from distributions with a high degree of correlation, which therefore exhibit a larger amount of multi-information.

3.4.1 *Coarse Graining*

Interestingly, this definition also gives the opportunity to build hierarchies by considering coarse to fine grained observers, which then leads to a decomposition of self-organization. If k-observers are grouped to one coarse-grained observer variable $\tilde{X}_i^{(t)}$, the multi-information term

$$I(\underbrace{X_{i_0}^{(t)}, \ldots, X_{i_1}^{(t)}}_{\tilde{X}_1^{(t)}}, \underbrace{X_{i_1+1}^{(t)}, \ldots, X_{i_2}^{(t)}}_{\tilde{X}_2^{(t)}}, \ldots, \underbrace{X_{i_{k-1}+1}^{(t)}, \ldots, X_{i_k}^{(t)}}_{\tilde{X}_k^{(t)}}) \tag{3.10}$$

can be decomposed into $k+1$ multi-information terms (see Friedman et al. (2006) and Polani (2008) for a derivation)

$$= I(\tilde{X}_1^{(t)}, ..., \tilde{X}_k^{(t)}) + \sum_{j=1}^{k} I(X_{i_1}^{(t)}, ..., X_{i_j}^{(t)}). \tag{3.11}$$

The decomposition now allows the separation of organization that is apparent within individual parts of the system, where each part is a coarse-grained observer, and organization, that can only be explained as an interaction between coarse-grained observers.

This allows to see whether there are parts of the system that dominate the process of self-organization. For example by grouping individual observers by common properties of the observed variables, it is possible to see whether a specific property has a higher contribution to the self-organization or whether most of the self-organization is a result of interaction between different coarse-grained observers.

It should be noted that coarse-graining allows easy regrouping of the variables, but it is not possible to simply recode the variables i.e. introducing a new set of observer variables $Y_i = f_i(X_1, ...X_n)$ and deduce a simple relation between the two multi-information terms $I(X_1, ..., X_n)$ and $I(Y_1, ...Y_n)$ (Polani, 2008).

3.4.2 *Choice of Observers*

In many cases there are natural choices for observers: multi-dimensional measurements where each measurement device is an observer variable, spatial structures that allow for a decomposition or multi-agent scenarios where the state of each agent acts as an observer. But sometimes there might be no obvious choice. By choosing observers that are maximally independent, the multi-information gets as small as possible, one might lose the appropriate 'perspective' to detect organization. These observers do however correspond to an independent component analysis and the concept of 'emergent descriptions' by Polani (2004,2006).

On the other hand, a maximization of multi-information in general does not lead to the desired outcome either. For n variables, one can simply observe the whole system n times to maximize the multi-information to $(n-1)H(X_1, ..., X_n)$ and organization would coincide with an increase of entropy. This is certainly not a desired outcome.

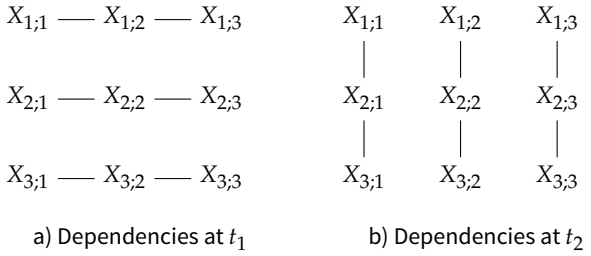


FIGURE 3.5 Showing the dependencies of random variables at different times for the observer choice example. Connected variables are copies of each other, unconnected variabled are independent.

One might be tempted to define maximal observers as those that maximize the increase of multi-information between time t and $t+T$ and subsequently call this increase the objective amount of organization. But the following example shows why this is doomed to fail. Let $X_{i;j}$ denote a matrix of binary random variables and $1 \leq i,j \leq 3$. Furthermore, the rows and columns as joint variables are denoted $X_{i;} = (X_{i;1}, X_{i;2}, X_{i;3})$ and $X_{;j} = (X_{1;j}, X_{2;j}, X_{3;j})$ respectively. Now, consider a random process consisting of such a matrix at every time t where at t_1 all variables in each row are copies of each other, but the rows themselves are independent of each other. On the other hand at time t_2 the dependencies are rotated by 90° and all variables in each column are copies of each other but the columns themselves are independent (see Figure 3.5). Using the rows and columns as observer variables the calculation of multi-information leads to

$$\begin{aligned} I(X_{1;}^{(t_1)}, X_{2;}^{(t_1)}, X_{3;}^{(t_1)}) &= 0 \text{ bit}, & I(X_{;1}^{(t_1)}, X_{;2}^{(t_1)}, X_{;3}^{(t_1)}) &= 2 \text{ bit}, \\ I(X_{1;}^{(t_2)}, X_{2;}^{(t_2)}, X_{3;}^{(t_2)}) &= 2 \text{ bit} \quad \text{and} & I(X_{;1}^{(t_2)}, X_{;2}^{(t_2)}, X_{;3}^{(t_2)}) &= 0 \text{ bit}. \end{aligned}$$

So depending on whether the rows or columns are observed the system seems to organize or actually loose organization. Even though, only a rotation (transposition of indices) is performed. Taking the individual variables as observers reveals that there has been no change in the organization of this system

$$\begin{aligned} I(X_{1;1}^{(t_1)}, \ldots, X_{3;3}^{(t_1)}) &= 6 \text{ bit}, \\ I(X_{1;1}^{(t_2)}, \ldots, X_{3;3}^{(t_2)}) &= 6 \text{ bit}. \end{aligned}$$

Removing the amount of observer induced organization via rotations and translations will be part of the next chapter, where I will revisit this problem in spatial systems. As Polani (2008) remarks, O-organization is not objective but depends very much on the perspective of the beholder, "but in a formally precise way" (Polani, 2008, p. 33). In practice this means that the choice of observers has to be done carefully and with the possible implications in mind.

3.4.3 *TSE Complexity*

It should be mentioned that multi-information is just one measure of a whole class of similar measures whose weighted average is called TSE-complexity (Tononi et al., 1994 and Ay et al., 2006a). For the definition I will use the index set notation with $V = \{1, \ldots, n\}$ and $X_V = (X_1, \ldots, X_n)$. The complexity measures are now defined as

$$C^{(k)}(X_V) := \frac{n}{k\binom{n}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) - H(X_V). \tag{3.12}$$

For $k = 1$ the multi-information $I(X_V) = I(X_1, \ldots, X_n)$ can be recovered, for $1 < k < n$ other complexity measures are the result. The averaged TSE-complexity is now defined as

$$C(X_V) := \sum_{k=1}^{n-1} \frac{k}{n} C^{(k)}(X_V). \tag{3.13}$$

These measures all share the desired property that they vanish if all X_k are either independent or deterministically related. The multi-information measures the deviation of the uncertainty of each variable to the overall mean stochastic dependence. Similarly $C^{(k)}(X_V)$ measures the deviation of the mean stochastic dependence of k-subsets of variables to the overall mean stochastic dependence. It can be shown that the $C^{(k)}$ are monotonic (Ay et al., 2011), i.e.

$$C^{(k)}(X_V) \leq C^{(k-1)}(X_V). \tag{3.14}$$

Using $C^{(k)}(X_V)$ for $k \geq 2$ or $C(X_V)$ as a measure of O-organization can be a viable alternative, especially considering the remarks by Ay et al. (2011) who shows that multi-information is maximized by pairwise interactions (for example the distribution of n-binary variables which are synced $p(111...1) = p(000...0) = 0.5$ displays only pairwise interactions), whereas the maximizers for TSE-complexity as well as $C^{(n-1)}(X_V)$ show interactions on all levels. Here I will continue using multi-information as the underlying measure of O-organization, firstly for practical reasons, including the availability of estimators for the computation of larger continuous systems and secondly, because it is not clear why higher order interactions necessarily have to make the system more complex than pairwise interactions. It seems that in biology complex systems use compartmentalization and build a hierarchy of lower order interactions to achieve higher complexity (Lehn, 2002 and Kolasa, 2005). The question whether this is due to information-theoretic or physical limitations would shed more light onto the question which measure is more suitable to measure organization.

3.4.4 *Interaction Complexity*

Speaking of higher and lower order interactions can be formalized: Kahle et al. (2009) construct interaction measures that give an interesting insight into what multi-information and TSE-complexity measure by offering an orthogonal view to them. The interaction measures are constructed using a hierarchy of interaction spaces (for the construction of these spaces, I refer the reader to (Ay et al., 2011))

$$\mathcal{E}_1 \subseteq \mathcal{E}_2 \subseteq ... \subseteq \mathcal{E}_n \subset \Delta(X_V) \tag{3.15}$$

where $\Delta(X_V)$ denotes the space of all probability distributions over X_V. Now it is possible to measure the Kullback-Leibler divergence between a probability distribution $P \in \Delta(X_V)$ and an interaction family $\mathcal{E}_i$ given by

$$D_{\mathrm{KL}}(P \| \mathcal{E}_i) := \inf_{Q \in \mathcal{E}_i} D_{\mathrm{KL}}(P \| Q). \tag{3.16}$$

The interaction measure can now be defined using the divergence

$$I^{(k)}(P) := D_{\mathrm{KL}}\left(P \,\|\, \mathcal{E}_{k-1}\right) - D_{\mathrm{KL}}\left(P \,\|\, \mathcal{E}_{k}\right). \tag{3.17}$$

This quantity measures the dependencies between k variables that cannot be explained by interactions of fewer variables. The interaction measures have an interesting relation to TSE-complexity as well as multi-information (Ay et al., 2011) as

$$I(X_V) = \sum_{k=2}^{n} I^{(k)}(P_{X_V}), \tag{3.18}$$

$$C(X_V) = \sum_{k=2}^{n} \frac{k(k-1)}{2} I^{(k)}(P_{X_V}). \tag{3.19}$$

Now, this also sheds a light onto the question why the maximizers of TSE-complexity show interactions on all levels, while the multi-information maximizers only show pairwise interactions. The coefficients for higher order interactions are increasing in the decomposition of TSE-complexity while the decomposition of multi-information stays flat (Ay et al., 2011). This means that keeping the total amount of TSE-complexity or multi-information fixed, there is a trade-off between contributions of different orders of interactions, which in case of TSE-complexity favors higher order interactions and in case of multi-information pair-wise interactions. Kahle et al. (2009) remark that, although $I^{(2)}$ measures only pairwise interactions in a closed system this does not mean that there cannot be higher-order correlations in an open or time evolving system that locally displays only pairwise interactions, i.e. $I^{(2)}$ is not the only positive of all interaction measures.

Thus, it is possible to get further insight into the structure of interactions, and multi-information thus also allows to distinguish different kinds of organization. Taking into account the remark by Bennett (1993) that self-organization obeys a law of slow growth it might be that there is a connection between the maximization of multi-information by locally pairwise interaction.

3.5 COMPARISON OF SC-ORGANIZATION AND O-ORGANIZATION

Both measures, SC-organization and O-organization, are good candidates to measure self-organization. The question of autonomy put aside, they both are defined as the increase of complexity measures. They have different areas of application and are "essentially 'orthogonal'" (Polani, 2008). SC-organization uses statistical complexity, based on the entropy of the causal state partition, whereas O-organization uses multi-information of observer variables as the underlying complexity measure.

SC-organization simply takes a time-series and without any prior knowledge tries to reconstruct the structure of the process. The obvious advantage of this is that there is no need to choose observer variables, as needs to be done when using O-organization. On the other hand it can be a disadvantage when dealing with spatial systems as the spatial structure

will be encoded in the global ϵ-machine, but not necessarily in a transparently accessible way (Shalizi, 2001). There is a spatial version of the ϵ-machine reconstruction algorithm, which can be used for all graph-like spatial structures using 'light-cones' instead of the one-dimensional past. Though, it seems not easy to use it for continuous environments. The focus of this measure is clearly on temporal dynamics, especially if there are no or no accessible compositional or spatial aspects, SC-organization is the better fitting device to employ.

O-organization on the other hand requires the choice of observer variables and focuses on spatial or compositional dynamics (Polani, 2008). In many cases there is a natural choice for observer variables as can be seen from the example as well from the next Chapter. From the properties associated with complex systems in (Ladyman et al., 2013) this measure reflects very well that an increase in complexity, hence organization, is expressed in correlations that are the result of interactions between parts of the system (spontaneous order). Furthermore, does the coarse graining allow for an easy investigation of hierarchical organization. If the system is continuous and not easily discretized, the measure of choice would be O-organization.

Ladyman et al. (2013) emphasize, that statistical complexity, though called 'complexity', is actually a measure of order. In particular statistical complexity is maximal for a periodic counter, that counts through all elements of $|\mathcal{X}|$ which is a system in perfect order. They conclude, that the picture of the one-humped curve (Figure 3.1) is misleading, and a true measure of complexity would assign such a system of perfect order a value of zero. As they proceed, a crystal has a perfectly ordered structure but lacks the adaptiveness, of for example a flock of birds, to be considered a complex system. However, as a measure of order in a noisy setting produced by a complex system, statistical complexity is considered a good candidate in (Ladyman et al., 2013). Thus adaptiveness is seen as a property of the complex system, but does not need to be a property of the complexity measure.

While I do believe that these are important considerations about complexity and that statistical complexity actually measures order in the presence of noise, however I would consider the crystal analogy in conjunction with adaptiveness possibly to vague in this context. It is not clear to me how the causal states of a continuous system with perfect spatial order and without adaptiveness would look like if the structure of the system is disturbed, for example the crystal is broken, as described in (Ladyman et al., 2013).

Moreover, it can be seen in the next chapter, that the O-organization of a spatially ordered system like a crystal is very low, unless local variations of individual particles are reflected elsewhere in the system. However, this is not an answer to the question of adaptiveness, and there might be adaptive structures that exhibit perfect order similar to a crystal, unless they interact for example with an obstacle where the structure adaptively recovers from the interaction. The question is now, if the complexity measure cannot distinguish both should it still be considered a complexity measure? If not, it seems to me that an interventional

measure of complexity is needed, one that measures the structure of the system under perturbations.

Before I give an overview of methods to estimate multi-information from empirical data to quantify O-organization, I will make a last remark that show another connection between both measures of self-organization: There is an interesting parallel between SC-organization and O-organization which supports the 'orthogonality' analogy remarked by Polani (2008). As mentioned earlier, statistical complexity is maximal for a periodic counter. This also means that for any process $\overleftrightarrow{X}$ it is possible to increase the complexity simply by adding a counting state and thus copying the original state space n times (where n is a prime). So instead of $\mathcal{X}$ the underlying state space is now $\mathcal{X} \times \{1, ..., n\}$ and the dynamics is changed so that in each time step the counter is increased by one, independent of the dynamics of the actual process, thus the causal states will also include this counter. This shows that statistical complexity requires that states are consistent over time, i.e. a random permutation of states (or better their labels, as not the individual samples are permuted in a different way, but only the labels of the states) in each time step would lead to zero statistical complexity (or the same value as the original system if only the counting state are permuted).

On the other hand, a permutation of the state labels in each time step leaves O-organization unchanged as multi-information among observers is independent of the dynamics of the process (unless observers are chosen that depend on the dynamics). A similar effect can be observed, however not in the dimension of time, but in the, possible spatial, dimension of the observers. Suppose, there are l observers $X_1^{(t)}, ..., X_l^{(t)}$ then the multi-information can simply be increased by making n coupled copies of the system and now considering $l \cdot n$ observers for this system. Here the implicit underlying assumption is not consistency of states over time, but consistency among observers, and thus often spatial consistency. Again if the state labels of all observers are randomly permuted, and thus there is no consistency among observers, the multi-information of the whole system will be zero (and the same value of the original system if the same permutation is applied among all observers of one copy). This demonstrates nicely where the idea of 'orthogonality' stems from.

3.6 ESTIMATION OF MULTI-INFORMATION

A practical advantage of O-organization is that it is quite easy to apply it to continuous processes. I will now give an overview over available estimation methods. There are several methods to estimate mutual-information, most of which can also be used to estimate multi-information. There are mainly four classes of estimators: Kernel based approaches (Moon et al., 1995, Steuer et al., 2002 and Suzuki et al., 2008), binning estimators (Hausser and Strimmer, 2009, Strong et al., 1996, Treves and Panzeri, 1995 and Panzeri and Treves, 1996), partition estimators (Fraser and Swinney, 1986 and Darbellay and Vajda, 1999) and estimators based on the Kozachenko-Leonenko entropy estimator (Kozachenko and Leonenko, 1987, Victor, 2000 and Kraskov et al., 2004) .

3.6.1 *Kernel Based Approaches*

Kernel density estimators can be used to estimate the probability density functions $p(x_1, ..., x_n)$ and $p(x_1)$,...,$p(x_n)$ involved in the calculation of multi-information (as done with mutual information in (Moon et al., 1995 and Steuer et al., 2002)). The probability density functions are estimated using Gaussian kernels

$$\hat{p}(\mathbf{x} = x_1, ..., x_n) = \frac{1}{m(2\pi h)^{d/2}} \sum_{i=1}^{m} \exp\left(-\frac{(\mathbf{x}^{(i)} - \mathbf{x})^\top (\mathbf{x}^{(i)} - \mathbf{x})}{2h^d}\right) \tag{3.20}$$

where d is the dimension of the space underlying the joint distribution $X_1, ..., X_n$, m the number of samples and $\mathbf{x}^{(i)}$ is the i-th vector valued sample. The marginals are estimated respectively and the kernel bandwidth is determined by Silverman's (1986) rule

$$h = \sigma \left(\frac{4}{m(d+2)}\right)^{\frac{1}{d+4}} \tag{3.21}$$

with σ being the average marginal standard deviation of the samples. The estimate is now obtained by numerical integration of the multi-information integral

$$\hat{I}(X_1, ..., X_n) = \int_{x_1,...,x_n} \hat{p}(x_1, ..., x_n) \log \frac{\hat{p}(x_1, ..., x_n)}{\hat{p}(x_1) \cdots \hat{p}(x_n)} dx_1 \cdots x_n \tag{3.22}$$

or as Steuer et al. (2002) noted by using the simplified formula

$$\hat{I}(X_1, ..., X_n) = \frac{1}{m} \sum_{x_1,...,x_n} \log \frac{\hat{p}(x_1, ..., x_n)}{\hat{p}(x_1) \cdots \hat{p}(x_n)} \tag{3.23}$$

as a cheap plugin for a numerical integration, which can be used if the samples were drawn independently.

However, directly approximating the pdfs involves the division of estimated quantities, which can be overcome by directly estimating the ratio in the logarithm. This approach, called MLMI (Maximum Likelihood Mutual Information), was introduced by Suzuki et al. (2008) for mutual information, but the approach is easily generalized for multi-information. It uses kernel density estimation via a maximum likelihood optimization on

$$w(x_1, ..., x_n) := \frac{p(x_1, ..., x_n)}{p(x_1) \cdots p(x_n)}. \tag{3.24}$$

The problem with this approach is that it involves convex optimization and bandwidth parameter selection via cross validation. In tests this resulted in an estimator several orders of magnitude slower than all the other estimators that I will introduce in the next sections. Therefore, I will only consider the pdf based kernel based estimator in my quantitative comparison.

3.6.2 *Binning Estimators*

Discrete entropy estimators can be used for multi-information estimation of continuous samples if the samples are quantized using a binning algorithm. A good review on histogram binning can be found in (Scott and Sain, 2005). There are several ways to determine the bins for continuous samples. Hausser and Strimmer (2009) use the Freedman-Diaconis inter quartile rule (Freedman and Diaconis, 1981) for estimating mutual information in gene regulatory networks while Slonim et al. (2005) and Lee et al. (2012) advocate adaptive approaches such as maximum entropy binning (Olsson et al., 2005) because fixed bins "break the coordinate invariance of mutual information" (Slonim et al., 2005, p. 2). Sturges' and Scott's rule (Scott and Sain, 2005) are also used often to determine bin widths.

Once the continuous samples are quantized, an entropy estimator $\hat{H}$ is used to estimate multi-information:

$$\hat{I}(X_1, ..., X_n) := \sum_{i=1}^{n} \hat{H}(X_i) - \hat{H}(X_1, ..., X_n). \tag{3.25}$$

The entropy estimator $\hat{H}(X)$ takes event counts c_i (that is the number of occurences of each event $x_i \in \mathcal{X}$ in the samples) as inputs. In the case of binned data, i denotes the index of the bin and X is the quantization of the continuous distribution according to the bins. Now, the most simple estimator is the plug-in estimator

$$\hat{H} := -\sum_{i=1}^{|\mathcal{X}|} \hat{p}_i \log \hat{p}_i \tag{3.26}$$

where $\hat{p}_i$ is a frequency estimate. In the case where the maximum likelihood estimator $\hat{p}_i^{ML} = \frac{c_i}{m}$ is used, m is the amount of observed samples. While $\hat{p}_i^{ML}$ is unbiased as a probability estimator, the derived plug-in entropy estimator $\hat{H}^{ML}$ however is biased. A first order bias correction is provided by the Miller-Meadow estimate

$$\hat{H}^{MM} := \hat{H}^{ML} + \frac{c_{>0} - 1}{2m} \tag{3.27}$$

where $c_{>0}$ is the amount of events with positive counts (Miller, 1955). There is also a whole class of Bayesian estimators that assume a prior distribution on X to estimate probabilities (Hausser and Strimmer, 2009). The NSB estimator (Nemenman et al., 2002) also uses a Bayesian approach, but with a prior that assumes a uniform distribution of all possible entropy values.

At last there are two other discrete estimators discussed in (Hausser and Strimmer, 2009). Firstly, the Chao-Shen Estimator H^{CS} which is a combination of the Horovitz-Thompson entropy estimator in combination with a corrected probability estimate (Vu et al., 2007). And secondly, introduced in (Hausser and Strimmer, 2009), the James-Stein Shrinkage Estimator $\hat{H}^{Shrink}$ which is a regularized Maximum-Likelihood estimator. I will not introduce these estimators in further detail here, a review of these discrete entropy estimators

can be found in (Hausser and Strimmer, 2009). The result of their comparison is that the maximum likelihood and Miller-Meadow estimators perform worst of all tested estimators on the considered discrete scenarios, while the Bayesian estimators give very mixed results depending on the data. The NSB, Chao-Shen and shrinkage estimators perform best, but the NSB estimator was slower than the shrinkage estimator by a factor of roughly 1000.

The main problem of these estimators is that in scenarios with only few samples their bias correction approaches do not work sufficiently (Victor, 2000). This is also true for the so-called *shuffled information* bias correction for mutual-information (Optican et al., 1991 and Chee-Orts and Optican, 1993) which generates a bootstrapped estimate of the joint distribution of the independent marginal variables which is then subtracted from the actual mutual information estimate. This method is easily extended to multi-information, however Panzeri and Treves (1996) show that in different scenarios this method may over- and underestimate the bias of the information estimation.

Another approach to estimation of entropy, mutual information as well as multi-information uses an extrapolation of estimates to achieve a bias correction (Strong et al., 1996 and Slonim et al., 2005). For mutual information it can be shown (Herzel and Groe, 1995 and Treves and Panzeri, 1995) that the estimate corrected by the first order term of the systematic error is given by

$$\hat{I}(X_1; X_2) \approx I(X_1; X_2) + \frac{B_{X_1X_2} - B_{X_1} - B_{X_2} + 1}{2m} \tag{3.28}$$

where $B_{X_1X_2}$ is the number of bins for the joint variable and B_{X_1} and B_{X_2} the number of bins for the individual variables respectively. In most cases $B_{X_1X_2} = B_{X_1}B_{X_2}$, however the joint variable could be binned independently (for example in the case of multi-information where the jont bin size quickly grows to computationally expensive sizes). Now the estimate is a linear function of $1/m$ and the estimate can be improved by a linear approximation of $I(X_1, X_2)$ using estimates with different numbers of samples. This approach is called the 'direct approach' to mutual information estimation. Slonim et al. (2005) use this approach to calculate the multi-information between three variables via the chain-rule and thus reducing multi-information estimation to several mutual information estimations. The optimal number of adaptive bins B^* (for all dimensions) in (Slonim et al., 2005) is then determined by increasing the number of bins to the largest number where the shuffled information is still zero.

3.6.3 *Partition estimators*

Partition estimators use a similar approach as binning estimators, however here a histogram is not created using bins, but via an adaptive partition of the joint variable space. One of the earliest mutual-information estimators by Fraser and Swinney (1986) uses such an adaptive partition. Another popular partitioning estimator was presented by Darbellay and Vajda (1999). It recursively splits cells of the partition, starting with one cell containing

all samples, into equally sized subcells. The splitting is then tested using a χ^2-test for the dependence of the new subcells and in case the cells are independent enough, the new splitting is added to the partition. This is repeated until there is no partition with at least two samples in it, that can still be split. Now the estimation of mutual information is calculated via the number of sample points in the cells. In theory these methods could be transferred to estimate multi-information, but especially in high dimensional spaces the required trees to store the partition will quickly exceed available memory limitations.

3.6.4 *Kraskov-Stögbauer-Grassberger Estimator*

The Kraskov-Stögbauer-Grassberger Estimator (KSG) (Kraskov et al., 2004) based on the Kozachenko-Leonenko entropy estimator (Kozachenko and Leonenko, 1987) can be used to estimate multi-information directly from continuous samples. First applications of the entropy estimator to mutual information estimation were done by Victor (2000). The KSG estimator is a direct method, as it does not include the calculation of intermediate entropy estimations. The estimate is based on a k-nearest-neighbor search. The estimator for m samples and n variables is given by

$$I(X_1, ..., X_n) \triangleq \psi(k) + (n-1)\psi(m) - \langle \psi(c_1) + \psi(c_2) + ... + \psi(c_n) \rangle_x \tag{3.29}$$

where ψ is the digamma function and the brackets denote the average taken over all samples. The c_i depend on the samples and are defined as follows: let $N_k(x)$ denote the k-th neighbor of the sample x in the set of all samples using the following metric

$$|x' - x| := \max_{i \in \{1,...,n\}} |{x_i}' - x_i|_2 \tag{3.30}$$

which is the maximum over the euclidean distance over the marginal samples. Now c_i is defined as

$$c_i = |\{x' \in \mathcal{X} : |{x'}_i - x_i|_2 < |N_k(x)_i - x_i|_2\}| - 1. \tag{3.31}$$

The idea is that a high correlation between the variables leads on average to a low count (at least $k-1$ by the definition of the norm used for k-th neighbour) of samples per variable that are closer to the sample itself, in the i-th variable, than the k-th neighbor over all variables, thus maximizing the estimator.

3.6.5 *Comparison of Estimators in Literature*

For the estimation of mutual-information there are several sources of estimator comparisons available (Victor, 2000, Kraskov et al., 2004, Hausser and Strimmer, 2009, Khan et al., 2007 and Papana and Kugiumtzis, 2008). The comparison performed in (Hausser and Strimmer, 2009) only deals with discrete distributions (the only exception is the shrinkage estimator which is used with the Freedman-Diaconis binning rule on continuous data).

In (Victor, 2000) it is suggested that estimators based on the Kozachenko-Leonenko estimate converge slower as a function of the number of samples, but exhibit a lower variance compared to binning estimators and are unbiased. Kraskov et al. (2004) remark that while adaptive binning is better than using equal sized bins, the general problem of binned estimators is the systematic error introduced by approximating the actual mutual information $I(X_1; X_2)$ with the value of the quantized mutual information. Furthermore, they compare the KSG-estimator with Darbellay and Vajda's (1999) adaptive partitioning estimator for different bivariate probability distributions and conclude that, while the latter is faster to compute, the KSG-estimator shows a smaller bias by approximately one order of magnitude. They also compare the KSG-estimator to kernel based approaches. Where Steuer et al. (2002) showed that kernel density estimators have smaller bias and smaller standard deviation compared to binning estimators using the direct method by Strong et al. (1996), Kraskov et al. (2004) criticize that in Steuer et al. (2002) the kernel bandwidth is chosen to large. Though Silverman's (1986) rule is recommended in literature to avoid statistical errors, they argue that the estimator thus is "insensitive to the finer details of the distribution" (Kraskov et al., 2004, p. 11).

Among the more recent comparisons of mutual information estimators are Suzuki et al. (2008), Khan et al. (2007), Lee et al. (2012) and Papana and Kugiumtzis (2008). Suzuki et al. (2008) compare the KDE method, the KSG estimator and the Edgeworth approximation (a polynomial estimator (Hulle, 2005)) with their MLMI approach and conclude that MLMI performs better or as good as the best of all the other estimators in all scenarios tested (linear, quadratic, and non-linear dependence as well as independence). They also remark that the Edgeworth approximation works only well if the underlying distribution is close to a bivariate normal distribution. However, as I already mentioned earlier, MLMI is several orders of magnitude slower than direct KDE approaches, which are already among the slower estimators, and will therefore be dismissed for further comparisons.

Khan et al. (2007) compare the KDE estimator, the KSG estimator, Edgeworth approximation and an adaptive partitioning approach (Cellucci et al., 2005) in several scenarios (linear, quadratic, periodic and chaotic) with different signal-to-noise ratios. They conclude that the KDE estimator is the best choice for samples from high noise settings and the KSG estimator is the best choice for samples from low noise settings where $m \approx 100$. For $m \approx 1000$ samples, KSG performs best (independent of the noise level). They also provide parameters for the kernel bandwidth and k in case of the KSG-estimator. Best results are attained in their scenarios with the choice of a bandwidth as introduced by Silverman (1986) and with $k = 3$.

Lee et al. (2012) compare a fixed sized binning estimator, the KDE estimator and Darbellay and Vajda's (1999) adaptive partitioning estimator on respiratory and blood gas concentration data. They conclude that the adaptive partitioning approach was the only one that could detect mutual-information in agreement with the experimental results. Papana and Kugiumtzis (2008) compare binning estimators (fixed width and adaptive) with the KDE

estimator, the KSG estimator and also Darbellay and Vajda's (1999) adaptive partitioning estimator. They compare them on data obtained from Monte-Carlo simulations of deterministic chaotic maps. Again the KSG estimator provides one of the best results and depends, as Kraskov et al. (2004) already remarked, the least on the choice of the parameter.

3.6.6 *Comparison of Selected Estimators*

In the literature review above it became clear that the KDE estimator as well as the KSG estimator are good, and often the best, choices for measuring mutual information in a small data scenario. However most of the comparison only looked at mutual-information between one dimensional variables. Only Kraskov et al. (2004), Slonim et al. (2005) and Lee et al. (2012) consider examples where the dimension of the joint distribution exceeds two, but none of them explore systems of higher dimensionality than three. Furthermore, there have been no quantitative comparisons of the Chao-Shen and the Shrinkage estimators with the KDE and KSG estimators.

Hence, I will compare several multi-information estimators here. The comparison will be between the kernel density estimator with Silverman's (1986) bandwidth rule as introduced above **(KDE)**, the Kraskov-Stögbauer-Grassberger estimator **(KSG)** (Kraskov et al., 2004), as well as the Chao-Shen **(CS)** estimator (Vu et al., 2007), the Shrinkage **(SH)** estimator (Hausser and Strimmer, 2009) and the direct approach using the chain-rule of multi-information **(DA)** (Slonim et al., 2005). The following comparisons are considered:

1. **Binning estimators:** As there is no quantitative analysis of **(CS)** and **(SH)** as multi-information estimators, I compare them to the direct approach **(DA)** (using four sample points for the linear extrapolation as in (Strong et al., 1996)). Fixed sized bins with the Freedman-Diaconis rule as well as adaptive binning is used for **(CS)** and **(SH)**. In the adaptive binning case the bin number is determined as the maximum number of bins where the shuffled information of the data still vanishes (for this 50 bootstrap samples are drawn and the value). Estimators are compared in a low dimensional ($d = 3$ dimensions over 3 variables) and high dimensional ($d = 10$ dimensions distributed over 7 variables) setting with $m = 500$ and $m = 10000$ samples.

2. **Values of k in higher dimensions:** For **(KSG)** the question is whether the proposed values for k are also valid for higher dimensions. Different values for $k = \{2, 3, 5, 10, 50\}$ are compared in a 10 dimensional setting with 500 samples.

3. **High dimensional comparison:** The **(KSG)** and **(KDE)** estimators are compared in even higher dimensional settings with $d = \{10, 25, 100, 200\}$ dimensions and $m = 500$ as well as $m = 1000$ samples. This comparison is not done with any of the binning estimators, which already foreshadows their bad performance in high dimensional settings.

The comparison is done on multivariate continuous samples drawn from distributions where the multi-information can be calculated analytically. These are the three test scenarios:

1. **Multivariate Gaussian (MG):** Samples are drawn from a multivariate Gaussian distribution $X_1, ..., X_n \sim \mathcal{N}(0, \Sigma)$. The multi-information is

$$I((X_1, ..., X_{l_1}), (X_{l_1+1}, ..., X_{l_2}), ..., (X_{l_{d-1}}, ..., X_n)) = \frac{1}{2} \log_2 \frac{\prod_{j=0}^{d-1} \left|\Sigma_{l_j, l_{j+1}}\right|}{|\Sigma|} \tag{3.32}$$

   where $l_0 = 1$, $l_d = n + 1$, $|\Sigma|$ the determinant of Σ and $\Sigma_{l_j, l_{j+1}}$ the block matrix of Σ spanning from (l_j, l_j) to $(l_{j+1} - 1, l_{j+1} - 1)$. In this scenario, the covariance matrix is simply

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \rho \\ \rho & \cdots & \cdots & \rho & 1 \end{pmatrix} \tag{3.33}$$

   where $\rho \in [0, 1)$ is a correlation parameter determining the correlation between all variables.

2. **Multivariate Gaussian, Random Covariance (RCMG):** This scenario is equivalent to the last one, except that $\Sigma = I + \frac{1}{n} A^\top A$ where A is a random matrix with coefficients in $[\rho, 1)$ (lower coefficients lead to a small value of multi-information and are thus not very interesting). This gives similar results as the previous one, but the correlations between the variables are not as regular.

3. **Uniform boxes (UBX):** In this scenario samples are drawn from a uniform distribution over a set of random hyper-rectangles in $[0, 1]^n$ (intersecting hypercubes result in a higher probability of a sample drawn from the intersection). Entropy can be calculated in this scenario as a sum of integrals over constant functions on the intersections and differences of all hyper-rectangles. Now multi-information can be simply calculated as the difference of entropies as usual. Several runs with different sets of random hyper-rectangles are performed.

Each scenario is used for each comparison and in each run m samples are drawn 50 times.

3.6.6.1 *Comparison Results*

I will now discuss the results of the comparison of the estimators starting with the results for the binning estimators:

1. **Binning estimators:** From Figure 3.6 and Figure 3.7 it can be seen that the only sensible estimates are attained in the low dimensional setting of the **(MG)** scenario among which the Shrinkage estimator with the Freedman-Diaconis binning rule leads to the best results. The maximum entropy binning interestingly leads to the worst estimations. The problem is that in the case of maximum entropy binning, the shuffled information quickly reaches a non zero value in high dimensional settings, therefore keeping the amount of bins low to reduce the bias of the estimator. This however leads to a larger error in the estimate and especially in the high dimensional setting there is no good intermediate number of bins leading to a trade-off between bias from using to many bins and error from using few bins that would outperform the Freedman-Diaconis rule. I did not continue to investigate this in further detail as the overall performance in the $d = 10$ dimensional setting was so poor, that binning estimators were not considered for the next two comparisons.

2. **Values of k in higher dimensions:** The proposed values for k in the available literature tend to be somewhere between 2 and 50 (Kraskov et al., 2004). Furthermore Kraskov et al. (2004) remarked that the result of the estimator is not very dependent on k. I was able to confirm this for a higher dimensional ($d = 10$) setting as the results in Figure 3.8 show. It can be seen that choosing a small k is preferable, especially in the case of close to Gaussian data. However, varying k in the range of 2 to 10 does not lead to drastic changes in the estimated information for a sample size of $m = 500$. What can be observed is an increasing bias with the actual amount of multi-information. There might be a possibility to account for this by using additional correction terms. For the purpose of this thesis though, it suffices that the estimation preserves the character of the mutual information, i.e. it is roughly monotonic with respect to the actual mutual information.

3. **High dimensional comparison:** The last comparison was conducted to compare the **KSG** estimator (with $k = 2$) and the **KDE** estimator in high dimensional scenarios ($d = 10, 20, 50$). It can be seen from Figure 3.9 that the **KSG** estimator provides better results in both the **MG** and **RCMG** scenario, while the result in the **UBX** scenario is not that clear. In the $d = 10$ dimensional case the **KSG** estimator correlates better with the actual mutual information, while for the $d = 20$ dimensional case both estimators seem to give fitting results, except for some outliers where the **KDE** estimator drasticly overestimates non-existing information. This seems to be a systematical problem of the **KDE** estimator: For small values of multi-information the errors of the marginal density estimations are first multiplied and then divided, which leads to an increase of the estimated value for distribution that have low multi-information. The difference between $m = 500$ or $m = 1000$ is only a fraction of the estimated value, in case of the **KSG** estimator the difference is within the errors of the 50 samples of estimates.

In summary, the **KSG** estimator is the only estimator which provided satisfactory results in the comparison and while still being biased can be used to detect the increase of multi-information in high-dimensional ($d \geq 20$) systems with only few ($m \approx 500$) samples available. As a final benchmark, I tested the **KSG** estimator in even higher dimensions ($d = 100, 200$) with different sample sizes (see Figure 3.10), the results show that already 100 samples give results that do not differ much from the estimates with a larger number of samples (taking the overall bias into account). Obviously the variance of the estimate decreases with a larger sample size. The second interesting result is that the bias scales almost linearly with the dimension. Thus it might be possible to add an additional bias correction term, at least for close to Gaussian data.

3.7 DISCUSSION

In this chapter, I introduced the concept of self-organization and presented several definitions and remarks about self-organization. I considered the definition of self-organization via statistical complexity by Crutchfield and Young (1989) and Shalizi (2001) and compared it to the alternative definition of self-organizatin via observers by Polani (2002). While these concepts are not new, the main contribution of the first part of this chapter is the literature review relating O-organization to other information-theoretic measures and the discussion about the observer choice, which will be continued for specific spatial systems in the next chapter.

The second part of this chapter consists of a comprehensive literature review of methods for the estimation of mutual information and concludes with a comparison of several estimators in the case of high-dimensional multi-information estimation. The results show that the Kraskov-Stögbauer-Grassberger Estimator (Kraskov et al., 2004) is the only suitable estimator for this task. An application of the results and methods discussed in this chapter can be found in the next chapter where multi-information estimation is employed to detect self-organization in spatial particle systems. In my initial search for good estimators of multi-information, I also tried some information geometric methods, which, to my knowledge have not been used for the estimation of mutual information or multi-information alike. The advantage of an information geometric formulation of mutual and multi-information is, that it is a one dimensional integral along a curve on a statistical manifold (Amari and Nagaoka, 2007). However, are the resulting terms not easier to estimate, on the contrary, the calculations get more complicated than the terms in the integral of the definition of continuous multi-information.

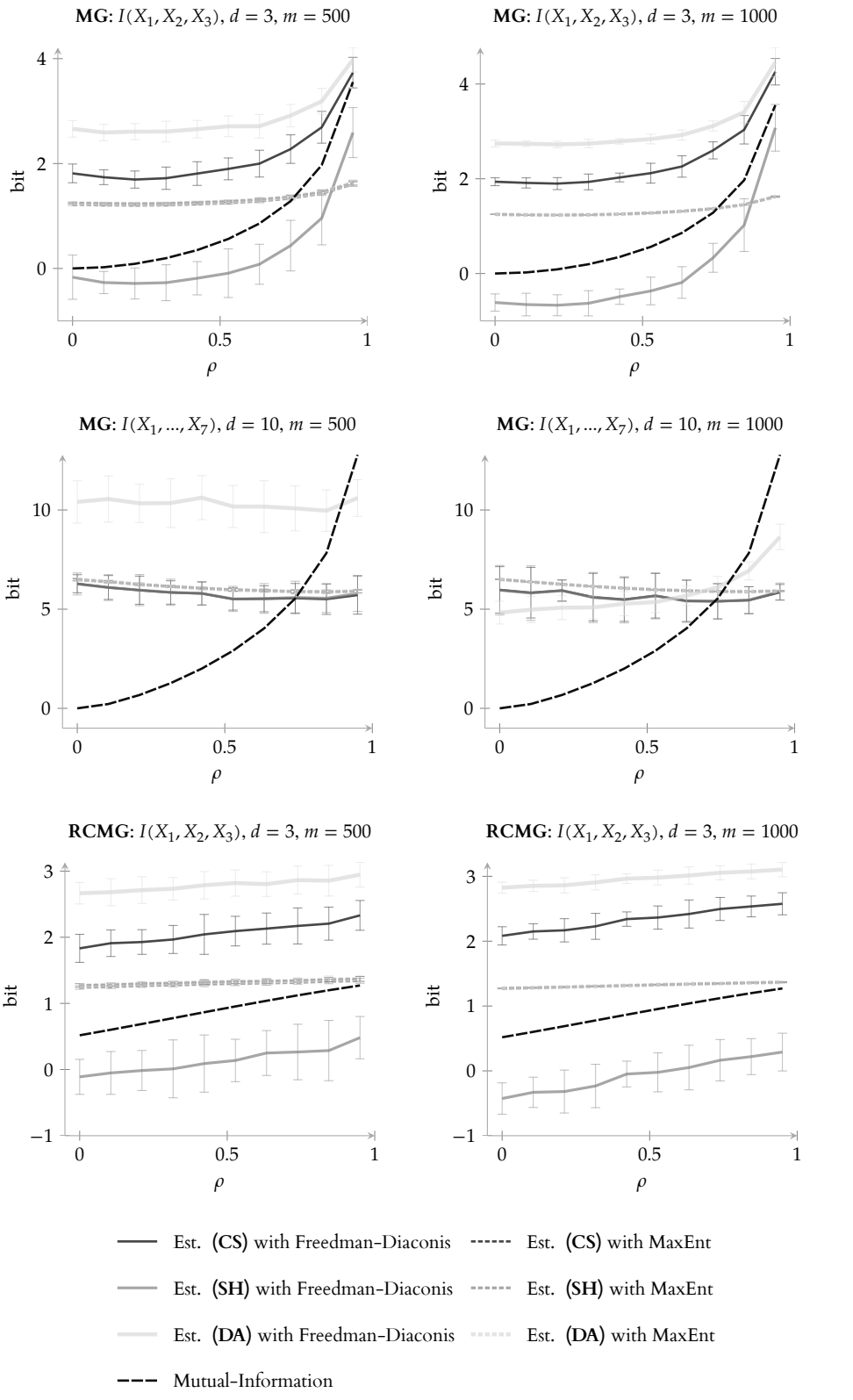


FIGURE 3.6 Comparison of different binning estimators and binning rules on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate) (part 1/2, see Figure 3.7 for part 2/2). Error bars denote one standard deviation from 50 estimation samples.

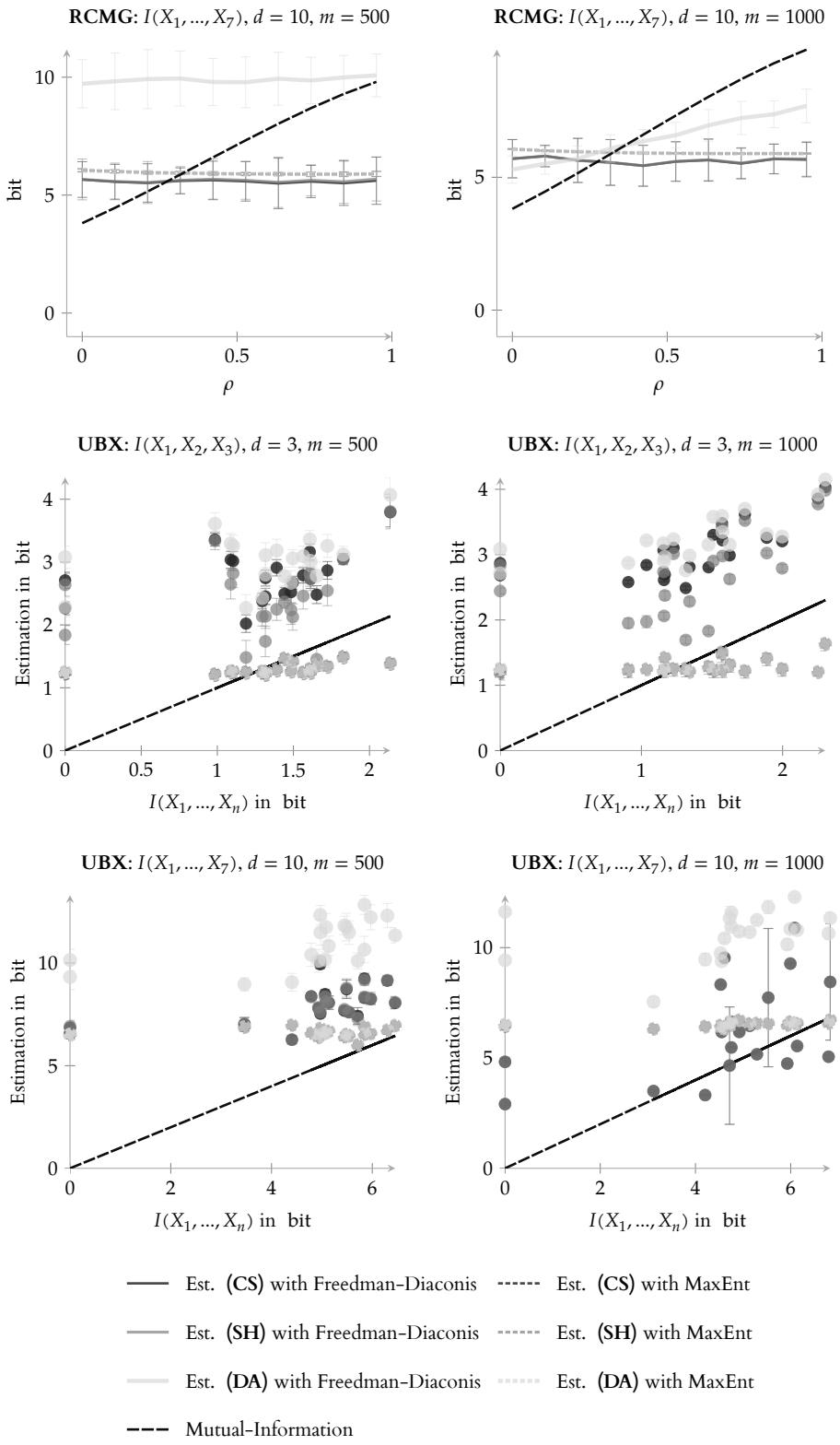


FIGURE 3.7 Comparison of different binning estimators and binning rules on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate) (part 2/2, see Figure 3.6 for part 1/2). Error bars denote one standard deviation from 50 estimation samples.

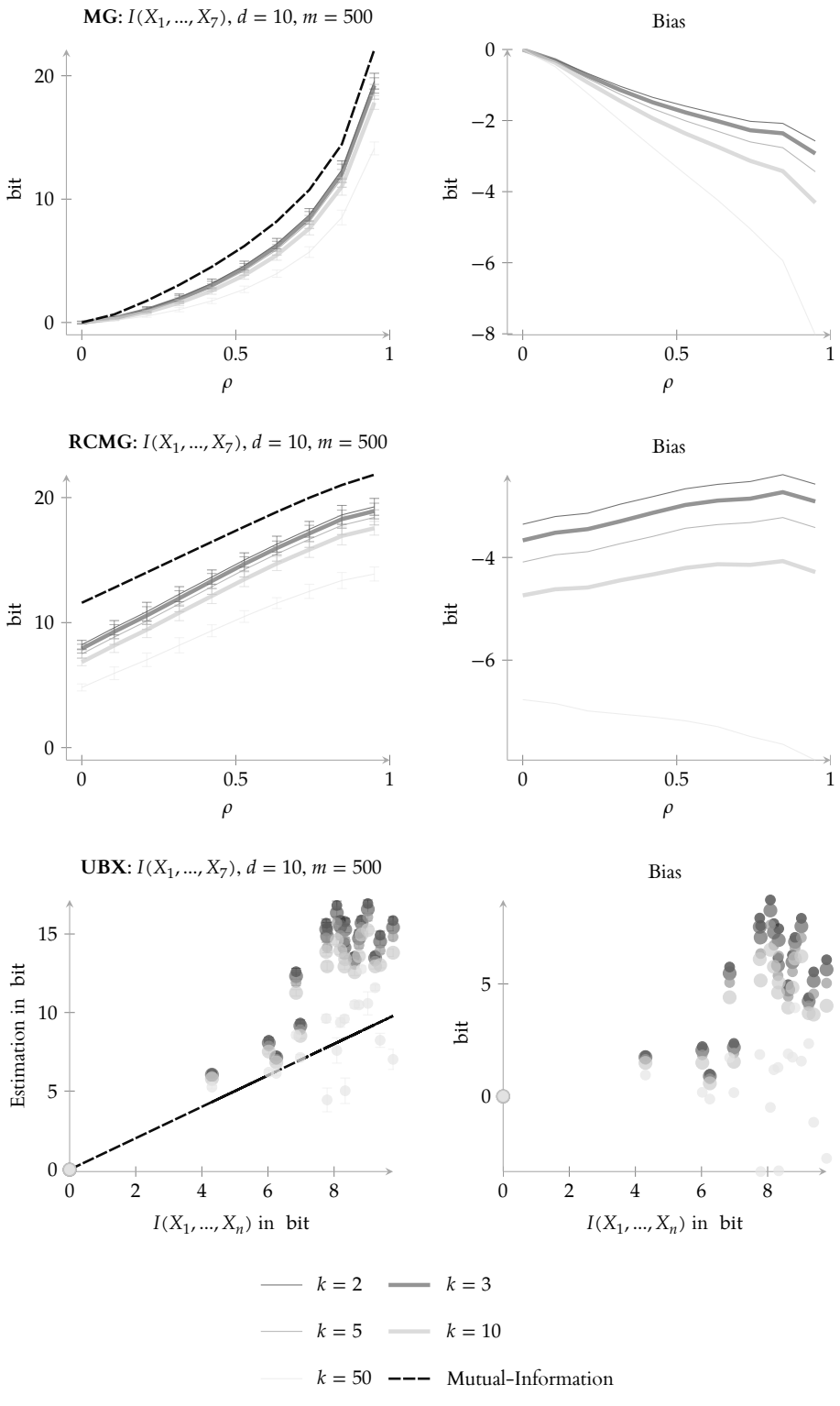


FIGURE 3.8 Comparison of different values of k (the k-th neighbour is used in the algorithm to estimate the multi-information) for the **KSG** estimator on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate) (part 2/2, see Figure 3.6 for part 1/2). Error bars denote one standard deviation from 50 estimation samples.

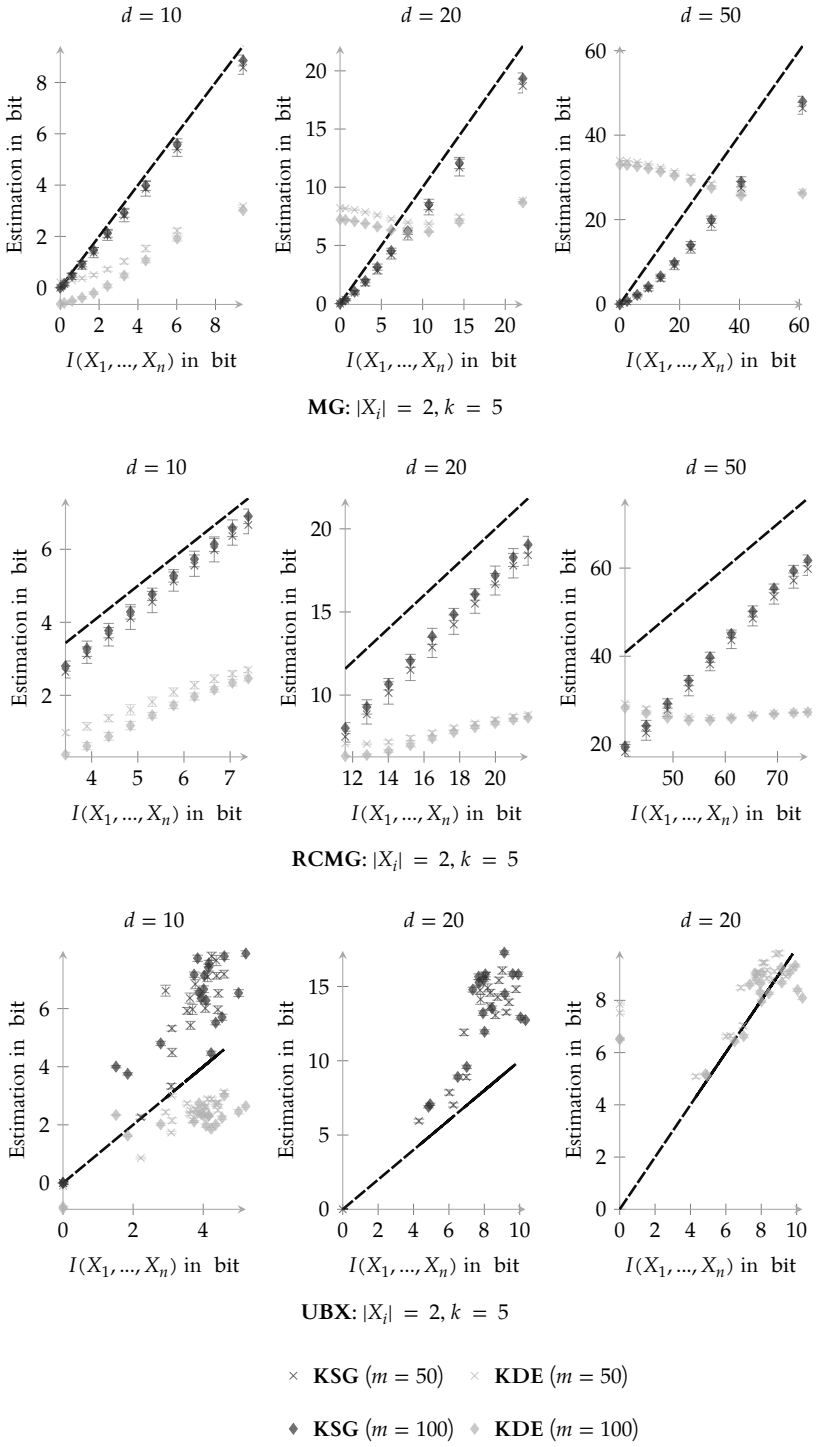


FIGURE 3.9 Comparison of **KSG** estimator and kernel density estimation on sample distributions (d denotes the dimension of the distribution, m the number of samples per estimate). The plot for $d = 50$ and the **UBX** scenario is missing as the analytical calculation of the mutual information exceeds memory limits in this case. Error bars denote one standard deviation from 50 estimation samples.

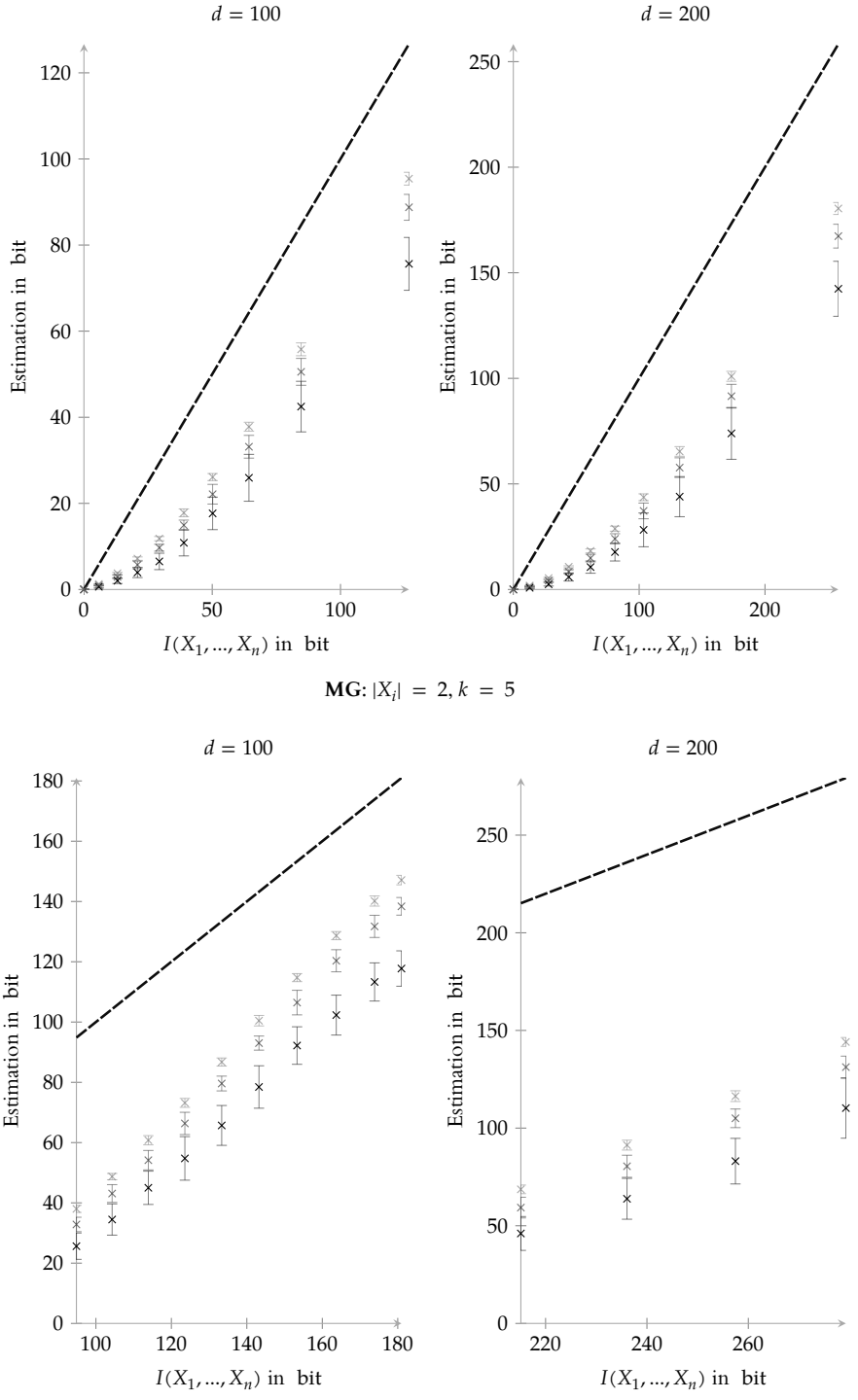


FIGURE 3.10 Comparison of estimates with different sample sizes ($m = 100, 500, 2500$) in high dimensional settings ($d = 100, 200$) using the **KSG** estimator. Error bars denote one standard deviation from 50 estimation samples.

SELF-ORGANIZATION OF PARTICLE SYSTEMS

4

» *It is not birth, marriage, or death, but gastrulation, which is truly the most important time in your life.* «
LEWIS WOLPERT, Unknown

The development of organisms is one of the most prominent examples of self-organization and the emergence of shapes. The process of forming shapes is usually an interplay between environmental dynamics (e.g. global physical rules), and agent actuations (e.g. a change of local properties) regulated through complex networks.

In the early stage of laying out body plans, morphological changes are induced mainly due to control of cell adhesion, cell motility and oriented cell division. In particular, differential cell adhesion prevents areas consisting of different tissues to mix and starts an automatic sorting process. This happens, if for example cells have been forced to mix in a solution (Wolpert et al., 2002). Gastrulation, the process of rearranging a ball of cells in the early stage of embryonic development into a more complex body structure, can be simulated by contractions in cell shape that then lead to an automatic rearrangement of cells forming an inner structure (Odell et al., 1980).

One important aspect of all these processes is that, in many cases, the information processing capabilities of the individual cells (i.e. agents) are severely limited, especially in scenarios that consider large collectives. In these cases the environmental dynamics dominate the process of organization while individual agents actively guide the process. Cells for example can change adhesion properties or partially contract. *Morphogenesis*, the formation of shapes, as will be seen, can be achieved purely by environmental dynamics up to certain limits. The process of shape formation can be seen as a selection of a configuration which fulfills certain properties. Thus, the course of a given process typically leads to a reduction of entropy. In the context of this thesis, I would like to reinterpret this as saying that there are information processing capabilities in the environment. This is justified by the view of the controlled dynamics of a system as an entropy reduction mechanism (Touchette and Lloyd, 2004).

These capabilities are often rooted in the structure of the space and the physical laws that govern it. Polani (2011) shows that consistency in the embodiment of agents reduces cognitive load, lack of such consistency increases it. Considering agents and environment as a joint information processing system, it follows that consistency or homogeneity of the space can also increase the information processing capabilities of the environment. A reduction of cognitive load for an agent means that the information needs to be processed elsewhere, one could say that embodied agents exploit the structure of the environment to process information. The information processing/entropy reduction capabilities that a system provides can also be used by non-reactive systems (for example, I consider particles here instead of autonomous agents). In particular they can be a driving force of self-organization.

In order to investigate the information processing capabilities of a morphogenetic process, I will use a model of particle collectives similar to the models by Doursat (2008b,2008a)

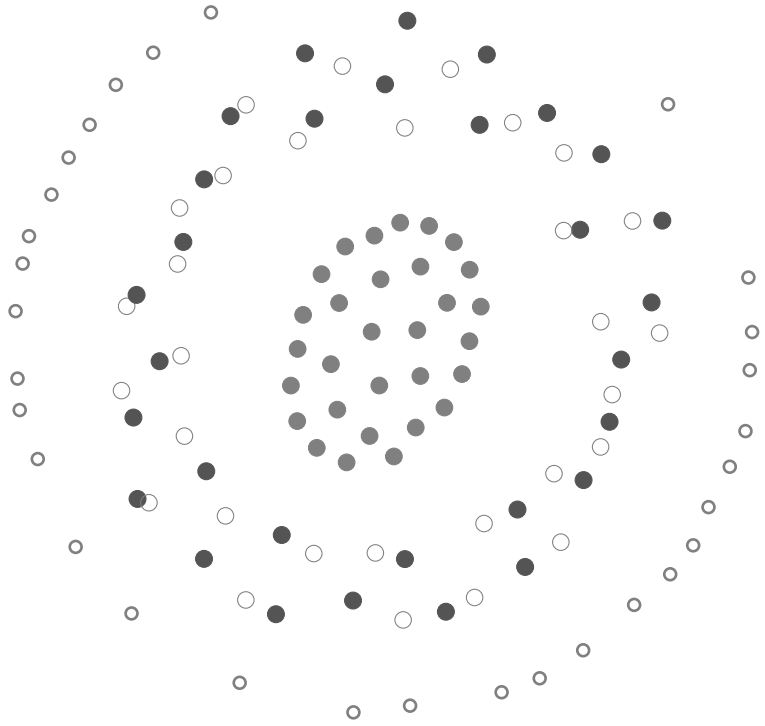


FIGURE 4.1 Example of a particle configuration.

and Sayama (2009) that mimics features of cell adhesion and cell motility to a certain grade. A human observer easily detects organizational patterns in simulation runs of this model. In many cases the resulting particle configurations even resemble the morphology of biological structures, showing features that look like membranes or nuclei, see Figure 4.1 for an example generated using the model I will introduce. However, a human observer is a quite subjective measure and not transferable. Using quantitative methods it is possible to investigate such a formation process in greater detail.

In self-organizing processes, individual parts of the whole system usually interact with each other, this is the case in particular in the particle model considered here. Interactions have been the basis for information-theoretic investigations before (Kahle et al., 2009), and can be closely linked to information storage and transfer (Wang et al., 2011, Lizier et al., 2008 and Lizier, 2011). A requirement that organization can occur is the spread of information through the system, which in turn requires interaction between individual parts of the system (Steudel and Ay, 2010).

4.1 PARTICLE COLLECTIVES & SELF-ORGANISATION

4.1.1 *The Particle Model*

There are numerous models of morphogenesis and pattern formation including reaction-diffusion models (Meinhardt, 1982), cellular automata (Wolfram, 1986), diffussion-lim-

ited aggregation (Witten Jr and Sander, 1981), L-systems (Prusinkiewicz, 1993) and agent based models (Bonabeau, 1997). The particle model I will use is based on the model by Doursat (2008b,2008a), and shares some similarities with the Swarm Chemistry model (Sayama, 2009). It mimics the way biological cells stick together by cell adhesion, allowing different types to recognize each other.

In the model, each particle interacts with all particles within a certain cut-off radius r_c. For reasons of simplicity, as well as to be able to have long range interactions, a cell-like tessellation, where interactions can only take place between direct neighbors of the tessellation, will be ignored as opposed to the model by Doursat (2008b). The equation of motion for each particle is given by

$$\dot{z}_i = \sum_{j \in N_{r_c}(i)} -F_{\alpha\beta}(|\Delta z_{ij}|_2)\Delta z_{ij} + w \tag{4.1}$$

where $\Delta z_{ij} = z_i - z_j$, $N_{r_c}(i)$ denotes the set of indices of particles within radius r_c of particle i and $F_{\alpha\beta}$ is a force-scaling function, α the type of particle i, β the type of particle j and w an additive white Gaussian noise term, where $w \sim \mathcal{N}(0, \sigma)$ with $\sigma \in [0, 0.1]$ throughout all experiments. The velocity is proportional to the force applied and thus the dynamics are studied in the strong limit of friction. This assumption holds for example for the motion of insects and cellular motility, in contrast to the movement of larger animals and humans which can build up momentum.

Now, the equation of motion can be solved using Euler-Maruyama integration (Kloeden et al., 1994 and Press et al., 1986). I used the following force-scaling function, similar to the model used in Doursat (2008b) (see Figure 4.2 for a function plot). In (Harder et al., 2011) two different force-scaling functions were considered. For the sake of clarity I only consider the first of the two functions from (Harder et al., 2011) here. The results with the second force-scaling function are comparable to the first function with a smaller cut-off parameter.

$$F_{\alpha\beta}(x) = k_{\alpha\beta}\left(1 - \frac{r_{\alpha\beta}}{x}\right) \tag{4.2}$$

The matrices $k_{\alpha\beta}, r_{\alpha\beta}$ define the interactions between the particles and have a strong impact on the dynamics of the experiment. Furthermore, there are two cut-off parameters r_b and r_c used in the simulation. The former limits the force that acts on particles that are very close to each other, such that if $|\Delta z_{ij}|_2 < r_b$ only a force of $F_{\alpha\beta}(r_b)$ is applied. The latter sets the applied force to zero if $|\Delta z_{ij}|_2 > r_c$, see Figure 4.2 for a plot of $F_{\alpha\beta}$ including the cut-offs. Values for the parameters were chosen from the following ranges: $k_{\alpha\beta} \in [0.0, 1.0]$, $r_{\alpha\beta} \in [0.25, 8.0]$ and $r_b = 0.25$ throughout all experiments. Choosing a non-symmetric matrix often leads to unstable dynamics or cycling patterns as in this case the preferred distance is mutually different, I therefore only consider symmetric matrices in what follows. The force-scaling function defines how much attraction or repulsion the particles show among each other depending on the type and distance between particles. For each type,

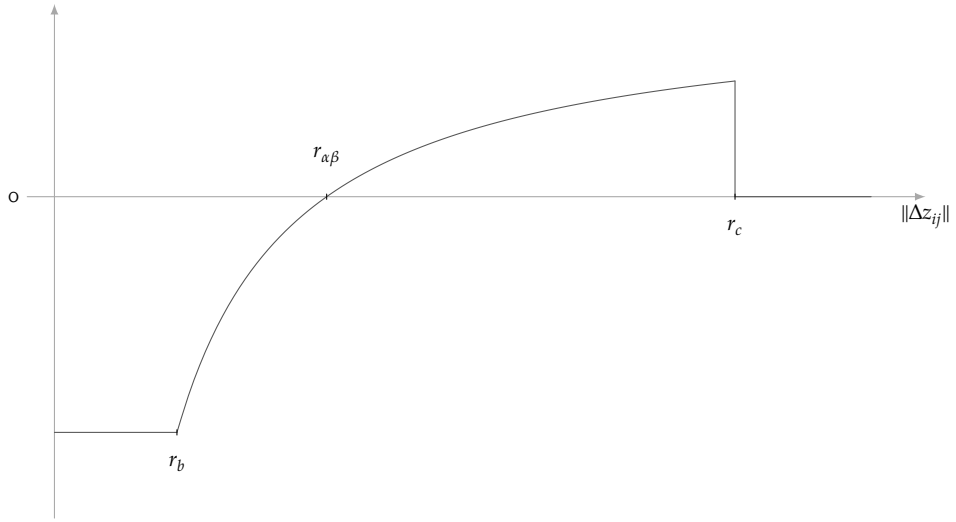


FIGURE 4.2 Plot of the force scaling function used for the particle dynamics, $r_{\alpha\beta}$ denotes the prefered distance between particles of type α and β. This radius can be directly specified as a parameter of the function. The long range attraction of is cut off by the radius r_c, whereas r_b limits the repellent force for particle that are very close.

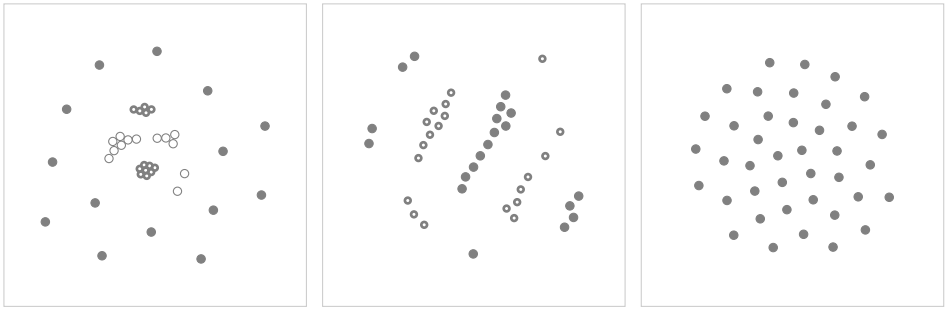


FIGURE 4.3 Examples of equilibrium states of particle collectives with different number of types.

the force-scaling function has a preferred distance of particles of other types, denoted $r_{\alpha\beta}$. By using smaller diagonal values than the off-diagonal elements in $k_{\alpha\beta}$ or $r_{\alpha\beta}$ it is possible to force clustering of particles of the same type.

In Figure 4.3 are three examples of equilibrium states of particle collectives. For the particle collective with only one type, a simple disc shaped pattern can be seen. The collective is considered to be in equilibrium, if for several time steps the sum of the L^2 norm of the sum of all forces acting on each particle is below a specific threshold.

4.1.2 *Measuring organization in particle collectives*

To measure self-organization within a particle collective using multi-information as introduced in Section 3.4 observer variables need to be defined. A natural choice would simply be the collection of variables denoting the positions of each individual particle. However, one needs to consider that certain transformations of the configuration leave the shape of the particle collective invariant. So, if these invariants are not considered, the measured

multi-information can be different from what I want to consider as organization towards a shape. But even if, in the stochastic limit, rotations and translations are equidistributed, factoring them out reduces the sparsity of samples in the space of possible configuration of particles in any case.

There are several accounts on spatial statistics and stochastics (Schinazi, 1999 and Liggett, 1985), however in these references interacting particle systems are defined as (continuous time) Markov processes on discrete domains while our experiments are in the continuous domain. In the area of geo-information systems and medical image processing, there is a large interest in statistical models of shapes, and there is a large body of literature on shape models (Davies et al., 2008, Small, 1996 and Dryden and Mardia, 1998). One particular problem, the alignment of overlapping images or shapes, is similar to the problem of reducing our experiment samples (i.e. the simulations) to an invariant representation.

Rotation, translation as well as permutation of particles of the same type leave the observable shape, as well as the dynamics involved, invariant. Let $ISO^+(2)$ denote the group of direct isometries (rotation, translation and identity) of the euclidean plane. This group now acts on the space of particle configurations $\mathcal{Z}$ by rigid body motions:

$$\mathcal{Z} \times ISO^+(2) \to \mathcal{Z}. \tag{4.3}$$

To account for permutations, let S_n denote the permutation group of n-elements, which also naturally acts on the space of samples by permuting the particle vectors for all time steps. Now it is possible to consider the subgroup $S_n^* \subset S_n$ that permutes only particles of the same type. The direct product $F = ISO^+(2) \times S_n^*$ then classifies all shape invariant transformations.

Note here that these transformations also have the property that they leave the dynamics of the system invariant. Let $z^{(t)} \in \mathcal{Z}$ denote the configuration of the particle collective at time t, then

$$p(z^{(t)}|z^{(t-1)}) = p(fz^{(t)}|fz^{(t-1)}) \text{ for all } f \in F \text{ and all } z^{(t)}, z^{(t-1)}. \tag{4.4}$$

This means that a configuration that is transformed will lead to a distribution of configurations in the future that is equivalent to the distribution of the transformed future states of the original configuration.

In the case that additionally the initial state is invariant under the action of this transformation group, that means $p(z^{(0)}) = p(fz^{(0)})$ for all $f \in F$, it follows that $p(z^{(t)}) = p(fz^{(t)})$ for all t and all $f \in F$. Thus it is easy to factor out the transformation group, and get random variables over the space of shapes (transformation invariant particle configurations). Factoring out all symmetries F from $\mathcal{Z}$ then leads to a reduced space of particle configurations $\mathcal{W}$ over which a random variable $W^{(t)}$ (the whole collective at time t) and corresponding observer variables $W_1^{(t)}, ..., W_n^{(t)}$ for a collective with n particles can be defined. Measuring multi-information on these derived random variables $W_1^{(t)}, ..., W_n^{(t)}$ now ignores certain

degrees of freedom, i.e., rotation, permutations of particles of the same type and translation. Now every configuration of particles z can be expressed as a permutation, translation and rotation of invariant coordinates w, i.e. for all z there exists w and $f \in F$ such that $z = fw$. Due to the group structure of F and the invariance of the states (at all times) under transformations of F

$$\begin{aligned} I(Z_1, ..., Z_n) &= \int_{\mathcal{Z}} p(z_1, ..., z_n) \log \frac{p(z_1, ..., z_n)}{p(z_1) \cdots p(z_n)} dz \\ &= \int_F \int_{\mathcal{W}} p(f(w_1, ..., w_n)) \log \frac{p(f(w_1, ..., w_n))}{p(fw_1) \cdots p(fw_n)} dw df \\ &= \int_F \int_{\mathcal{W}} p(w_1, ..., w_n) \log \frac{p(w_1, ..., w_n)}{p(w_1) \cdots p(w_n)} dw df \\ &= I(W_1, ..., W_n). \end{aligned}$$

Therefore, factoring out the transformation group F does not change the multi-information of the observers, in the case of an invariant system. I use an initial distribution of particles, which is uniform within a certain radius around the origin, so that particles are initially placed uniformly and independently on a centered disc. This initial distribution is still invariant with respect to rotation and permutation, but not translation invariant. However, the multi-information is generally invariant under transformation of homeomorphisms (Kraskov et al., 2004), therefore the above equality holds at all time as long as all elements of $f \in F$ are homeomorphisms of $\mathcal{Z}$, which is the case here.

4.1.2.1 *Indistinguishable particles*

If I make predictive statements about particles it is required that particles can be identified through time, otherwise the statistics about the future of a particular particle are skewed. That the interchangeability of variables has an impact on information processing and measurements has been considered before in terms of recoding equivalence (Crutchfield, 1990). By reordering the particles, the information to identify the same particle over time is lost and they become indistinguishable. To measure self-organization of shapes indistinguishable particles (if they have the same type) are desired and therefore I introduced the permutation group S_n^* as one set of shape invariant transformations. Distinguishing them would mean that there can be an event that increases the measurement of self-organization, but is not reflected in the shape and structure of the particle configuration. For example there could be a permutation of two particles of the same type that is always reflected by a permutation of two particles of same type elsewhere in the system. This would then be taken into account by the multi-information, but has no impact on the shape that is formed.

On the other hand I do not want to equate particles which have a different type, and show different interactions. Particles of different types should be distinguishable as permutations

of particles of different type would change the shape of the configuration. Additionally, if particles of different type would be indistinguishable this the particle dynamics would not be invariant to permutation anymore

The problem of indistinguishable particles and the related change of entropy is also a problem in thermodynamics where it is known as Gibbs phenomenon and Mixing paradox (Gibbs, 1874 and Jaynes, 1992). Only making a distinction between particles that show observably different behavior also agrees with the solution to this problem in physics (Jaynes, 1992). This is a very subtle problem, as the value of thermodynamic entropy now depends on how well the observer can distinguish them and the same is true for the value of multi-information. However, here I am measuring multi-information in a model, where it is known from construction which particles are distinguishable.

4.2 METHODS

This section describes how I derived estimates of multi-information from simulation samples of particle dynamics. Each simulation runs with a fixed number of n particles, l different types and each particle gets a fixed type assigned at the start of the simulation. The particles are located in the infinite two-dimensional plane $\mathbb{R}^2$ and are initialized with a uniform distribution on a disc of fixed radius. Each particle is of a specific type. The types can vary between different experiments, but the properties ($r_{\alpha\beta}$, etc.) of each type are fixed for all simulation samples of one experiment. The assignment of a type to a particle is fixed over the time of the simulation run. Each simulation run is a sample and is denoted by

$$\bar{z} = (z^{(1)}, ..., z^{(t_{\max})}) \tag{4.5}$$

where each time step is a vector of particle coordinates

$$z^{(t)} = (z_1^{(t)}, ..., z_n^{(t)}). \tag{4.6}$$

To gather statistics for an experiment, the simulation needs to run multiple times. The collection of all m samples is denoted

$$\mathfrak{z} = (\bar{z}_1, ..., \bar{z}_m)^\top = (\mathfrak{z}^{(1)}, ..., \mathfrak{z}^{(t_{\max})}). \tag{4.7}$$

Now, let the space of all particle vectors $z = (z_1, ..., z_n)$ be denoted $\mathcal{Z}$ and $Z^{(t)}$ the random variable over $\mathcal{Z}$ at time step t, so all $z^{(t)} \in \mathfrak{z}^{(t)}$ are samples of $Z^{(t)}$.

4.2.1 *Factoring out symmetries*

Next step is factoring out the symmetries for each time step as introduced in Section 4.1.2. The samples $z^{(t)} \in \mathfrak{z}^{(t)}$ for each time step t, the raw output of the simulations, are still with respect to a common coordinate system. I proceed by factoring out translations, rotations and permutations resulting in processed samples $w^{(t)} \in \mathfrak{w}^{(t)}$ for each time step t. In practice

this is done by expressing all particle configuration samples $z^{(t)} \in \mathfrak{z}^{(t)}$ with respect to its centroid. This is followed by aligning all configuration samples $z^{(t)}$ for each time step using an ICP (Iterative Closest Point) algorithm (Zhang, 1992 and Rusu and Cousins, 2011). The ICP algorithm, associates mutual points of two sets and minimizes the squared mean distance between the associated points by a linear transformation of all points of one set. This is iterated several times, where the associations are recreated after each minimization, so that each iteration gives a refinement of the associations and transformation.

For the application of the alignment the particle configuration is transferred to a three dimensional representation where the third coordinate of each particle is represented by its type, where the type coordinates are scaled by a factor a magnitude larger than the diameter of the collective. Thus the alignment respects the type of the particles. After the alignment the coordinates of all particle are reordered by types and correspondences. Correspondences between particles of different samples, but of the same type, are found using a nearest neighbor search within the ICP algorithm (implementation from the point cloud library (Rusu and Cousins, 2011)). This means that particles close to each other in different samples at the same time are considered to represent the same particle. Note that the notion of same particle establishes a correspondence between different samples at a specific time step. The correspondence between particles of the same sample, but different time steps is, however, lost in this process.

Equipped with this preprocessing, an isometry- and permutation-reduced representation of the particle collective is reached in terms of processed samples $w^{(t)} \in \mathfrak{w}^{(t)}$. I can now use the statistics of these samples to calculate the multi-information $I(W_1^{(t)}, \ldots, W_n^{(t)})$.

The invariant representation also has the advantage that the samples are much denser in the space of possible configurations which improves the quality of the estimates. It is important to note, that for statistics that need to track particles over time, one cannot use the permutation-reduced representation because one would lose any correspondence of particles over time, e.g. (Kondor, 2008).

4.2.2 *Estimation of multi-information*

To estimate the multi-information I used the Kraskov-Stögbauer-Grassberger Estimator (Kraskov et al., 2004) as introduced and compared to other approaches in Section 3.6. The estimate is based on a k-nearest-neighbor search. As the results of the comparison and literature review suggest a rather small value, $k = 5$ was chosen for all experiments. In my experiments the sample sizes vary from 500 to 1000. The comparison in Section 3.6 showed that the estimator is surprisingly good especially for such a sparse sampling in a high dimensional setting, even though there is a clear bias that increases with the amount of actual multi-information. Therefore the estimated multi-information is possibly less than the actual multi-information, however the estimator does capture increases and decreases of multi-information.

For large collectives, the alignment of samples and the estimation of the multi-information can still be a computationally expensive task. However, it is possible to reduce the dimensionality of the problem by introducing mean random variables. To do this a k-means clustering on the set of particles of each type can be performed and thus $l \cdot k$ mean variables $\hat{W}_1^{(t)}, ..., \hat{W}_{lk}^{(t)}$ are recovered, where l is the number of types. Now taking $I(\hat{W}_1^{(t)}, ..., \hat{W}_{lk}^{(t)})$ as an approximation measure for the multi-information $I(W_1^{(t)}, ..., W_n^{(t)})$ reduces the computation time. This must be done *carefully*, because the clustering process itself can introduce structure into the collective of particles, and thus can lead to a higher measurement of multi-information than actually is present. On the other hand, the clustering ignores all small scale self-organization processes, and hence the measured multi-information is less than the actual value. The experiments performed here are not using this clustering method.

4.3 EXAMPLES

One of the most simple examples is a particle system consisting of a single type (with a preferred distance between particles of $r_{\alpha\alpha}$) and three particles. The particles in this system move towards three points of mutual distance $r_{\alpha\alpha}$. In (Harder et al., 2011), we only looked at dynamics with a fixed noise of $\sigma = 0.05$. Here, I want to take a look at the three particle system first without noise ($\sigma = 0$). In this case, the equilibrium distribution is a Dirac delta function (where the three particles form an equilateral triangle in the configuration space) and thus the multi-information vanishes. Plotting the multi-information at each time step shows however, that there is an initial phase where the system organizes, followed

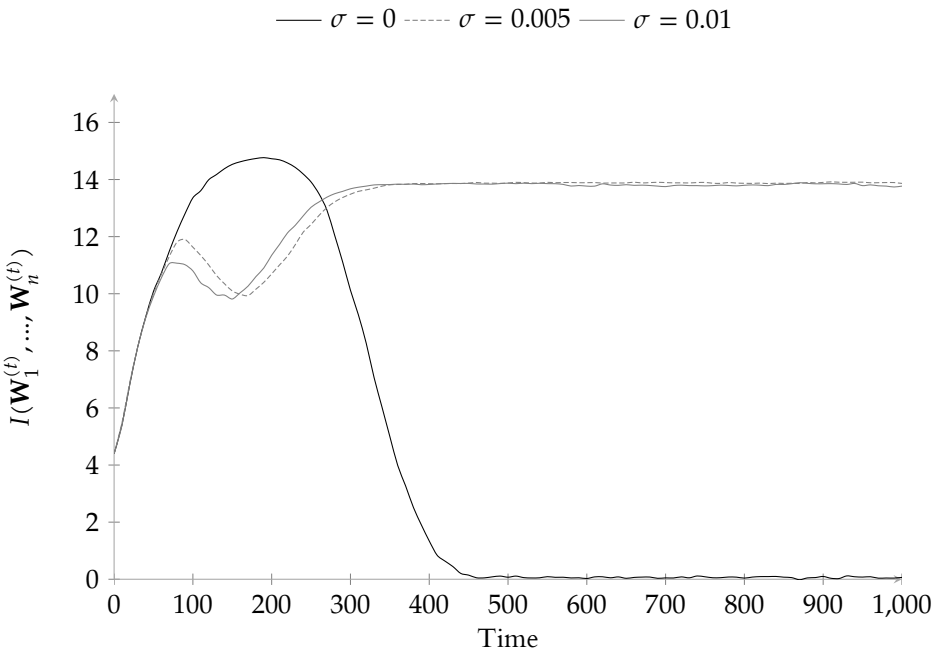
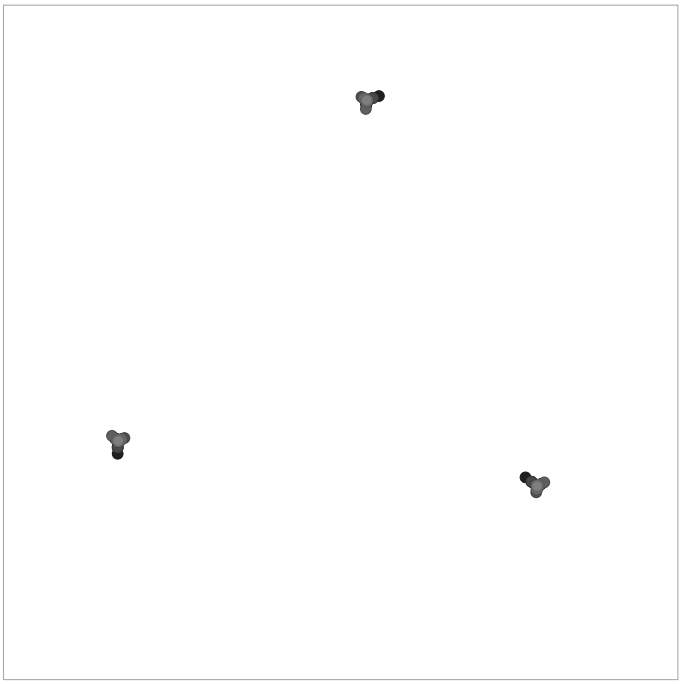
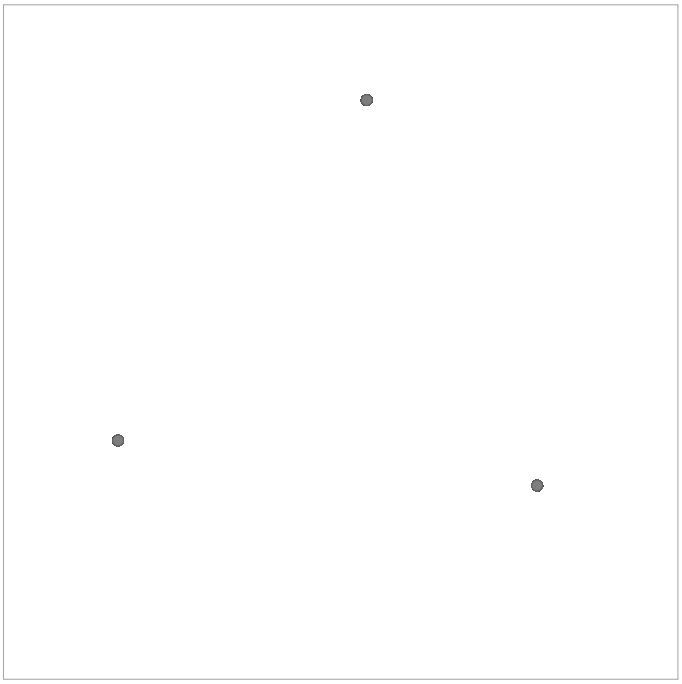


FIGURE 4.4 Multi-information between three particles of the same type for different noise levels ($m = 1000$ samples, $r_c = 10, r_{\alpha\alpha} = 2.5$).



a) Particle configuration samples at $t = 200$



b) Particle configuration samples at $t = 1000$

FIGURE 4.5 Plot of the samples from the noise free three particle example at two different time steps. The particle configurations are shaded by sample. Therefore, it can be seen in a) that the outliers along the three axes belong to the same samples. This is a sign for correlation between the particles and hints towards a larger amount of multi-information.

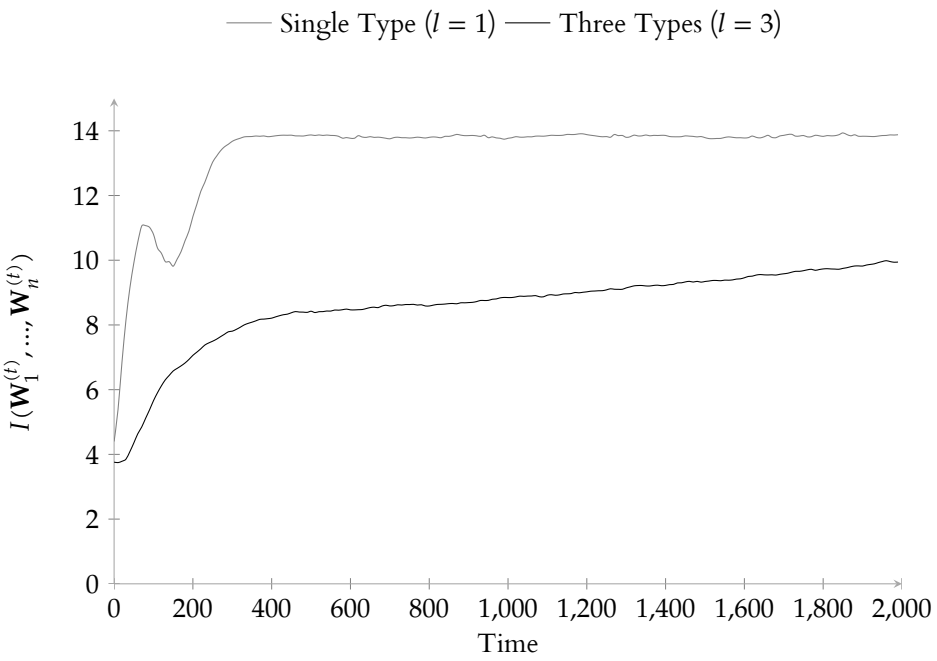


FIGURE 4.6 Multi-information between three particles with $l = 1$ and $l = 3$ types ($m = 1000$ samples, $r_c = 10$, $r_{\alpha\beta} = [[2.5, 0.1, 5], [0.1, 1, 0.5], [5, 0.5, 2]]$ and $k_{\alpha\beta} = [[0.05, 0.5, 0.1], [0.5, 0.2, 0.5], [0.1, 0.5, 0.3]]$).

by a 'cooling down' period where the system settles to the equilibrium configuration (see Figure 4.4).

In Figure 4.5 all samples of the noise free system are shown in an overlay plot. It can be seen that there is no variation in the samples at $t = 1000$, whereas at $t = 200$ the particle configuration can still vary with respect to the mutual distance of particles, though in a very constrained way.

If the dynamics are noisy ($\sigma = 0.005$, $\sigma = 0.01$), the cooling down period does not happen and the equilibrium is attained at a level of maximal organization over the evolution of the system (see Figure 4.4). There is a small dip in the initial phase, but it is not clear from a visual inspection where this short phase of cooling down comes from. Here, also the entropy estimates of the whole particle system and the sum of entropies of individual particles do not provide an explanation, as the estimators operate on spaces of different dimensions with different biases and are therfore hardly comparable (which is also the reason why a special estimator like the KSG estimator was needed in the first place).

What this initial example emphasizes is that a stochastic dynamical system needs a source of entropy to self-organize. In the case of deterministic dynamics, the source of entropy is the uniform initial disitribution (on a disc of radius 5), but the dynamics decrease the individual entropies until they vanish, in which case the multi-information is also zero and therefore it looks as if no self-organization occured. In the three-particle system with noise in the dynamics, the noise is reflected not only in the sum of individual entropies

but also in the overall systems entropy, but here correlated via the interaction between the particles and thus the multi-information does not drop. This is similar to the measure of information flow (Ay and Polani, 2008), where a causal influence between random variables is measured via the intervention at variables in a CBN. In this case, however, there are no direct interventions, only noise is added to the dynamics. In practice, measurements of the physical system are usually noisy, but in the case of quantized data or simulated systems one needs to be careful as the system might appear as if it disorganizes while the correlations between individual parts of the system cannot expressed. This happens due to missing noise in the system, which 'probes' the mechanisms that cause the correlation. Hence, in all experiments that follow an additive Gaussian noise with $\sigma = 0.01$ is used.

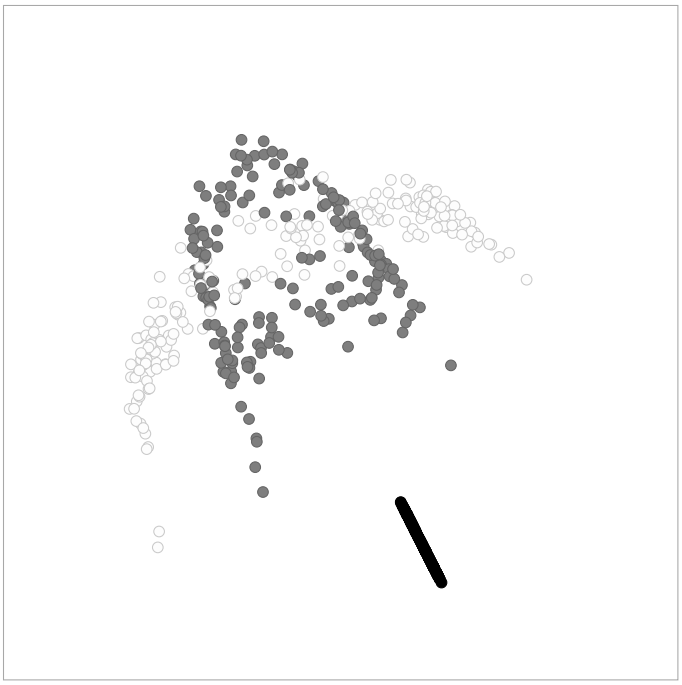
4.3.1 *Multiple Types*

Having only one type of particles is quite a severe limitation on the possibe interaction of particles. With three particles and three different types the configurations show more variation in shape (see Figure 4.7) though the actual organization (increase of multi-information) is less than the organization achieved by a system where all particles are of the same type (see Figure 4.6) suggesting that the varations in shape are not that correlated and the individual particles have a higher degree of freedom.

Most of the interesting self-organizing systems, however, do consist of more than three parts, especially living organisms, which consist of a large number of cells. Hence, I will now show an example of a systems with $n = 70$ particles and $l = 3$ types. The first observation is that this method can be used in practice to detect self-organization of high dimensional systems and there is a visual correlation between the formation process and the increase of the multi-information estimate as depicted for the $l = 3$ types example in Figure 4.8. In the beginning the sum of the marginal entropies $H(W_i^{(t)})$ is as large as the overall entropy of the system because there is no correlation between particles at all (this is not measured, however at $t = 0$ the entropies can still be calculated analytically). Over time, the marginal entropies decrease, however the overall entropy decreases even faster as the variations of individual particles are correlated. This then leads to an increase of multi-information over time. In Figure 4.8 it can also be seen that the final shapes show a certain variety, and there are two visually distinguishable categories of shapes which are shown at different time steps of the simulation runs.

4.4 RESULTS

After presenting some simple examples, I want to show first what happens when the number of particles is increased. In Figure 4.9 it can be seen that, in case of the three type system, the organization is increasing almost monotonically with the number of particles. This effect cannot be observed for the system with only a single type, where the organization is increasing less in comparison with the three particle system and interestingly decreases



a) Particle configuration samples at $t = 150$

b) Particle configuration samples at $t = 2000$

FIGURE 4.7 Plot of all samples of particle configurations of the three particle and three types example at different time steps. Each shade denotes a different type, all samples of configurations of three particles of three different types are overlayed in this plot.

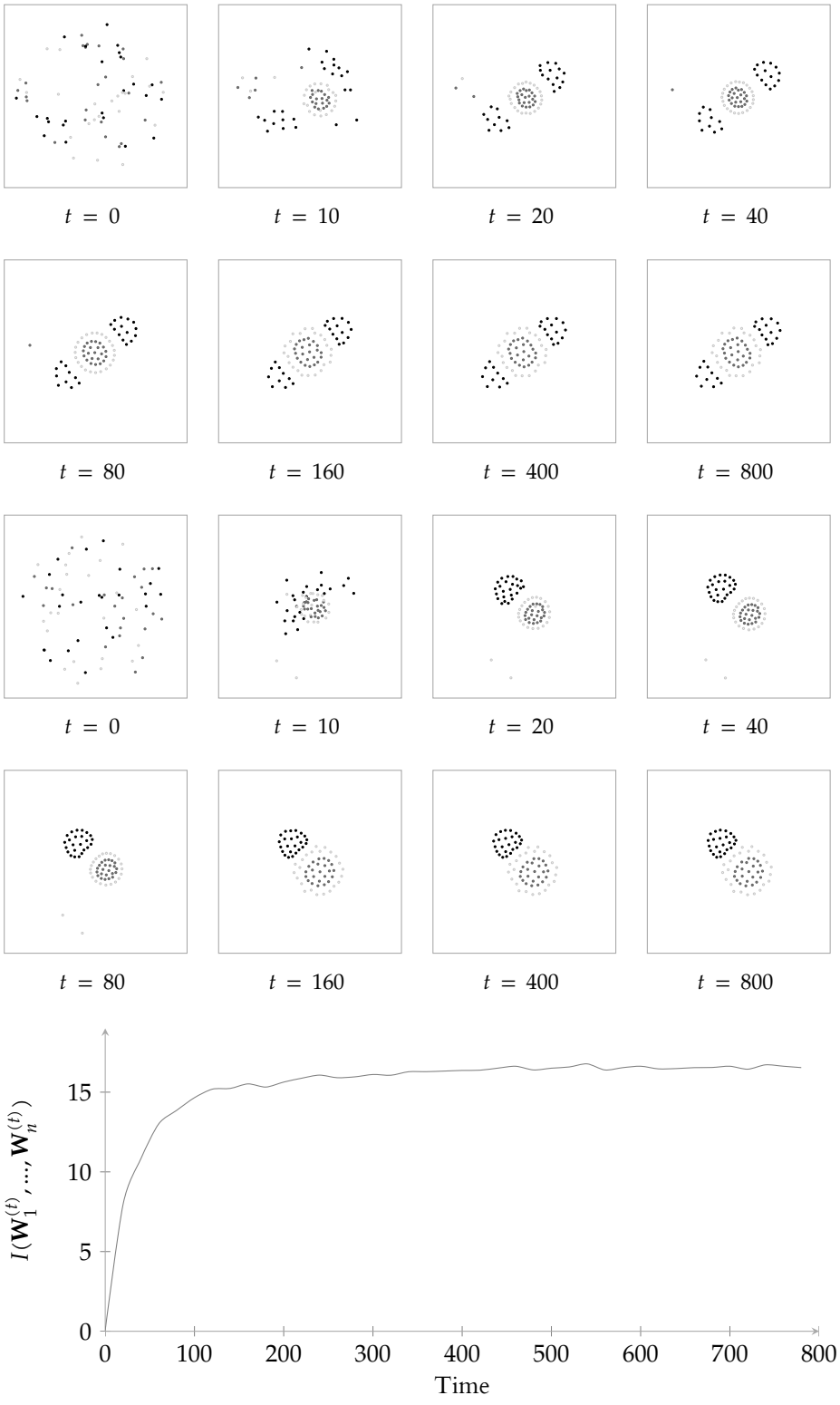


FIGURE 4.8 Multi-information between particles plotted against time with $n = 70$, $l = 3$, $r_c = 6.0$, $r_{\alpha\beta} = [[2.5, 5, 4], [5, 2.5, 2], [4, 2, 3.5]]$ and $k_{\alpha\beta} = [[0.6, 0.1, 0.1], [0.1, 0.6, 0.6], [0.1, 0.6, 0.6]]$ ($m = 500$ samples). The increase of multi-information correlates with the visual organization shown by snapshots of two samples at different times.

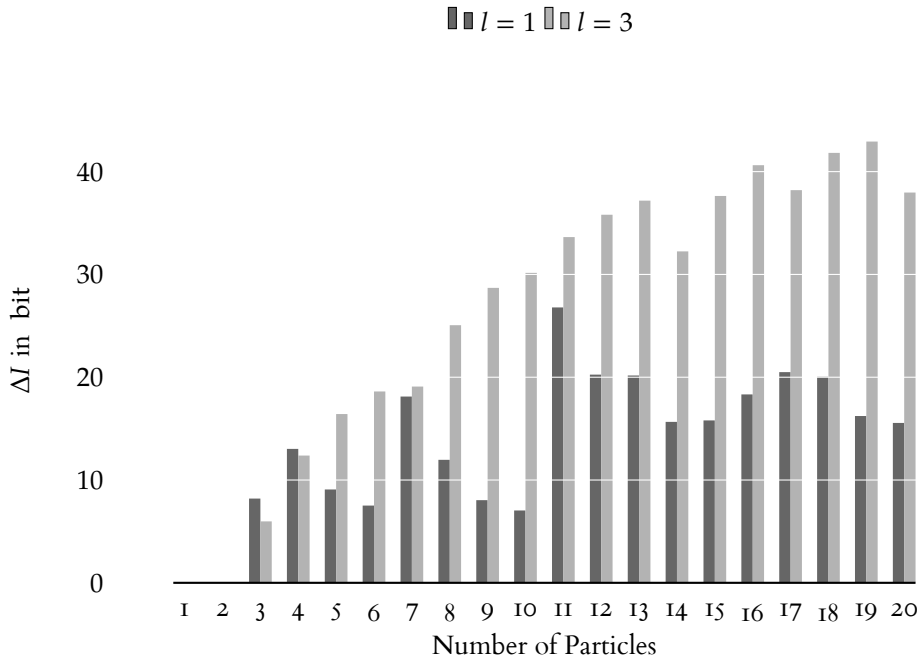
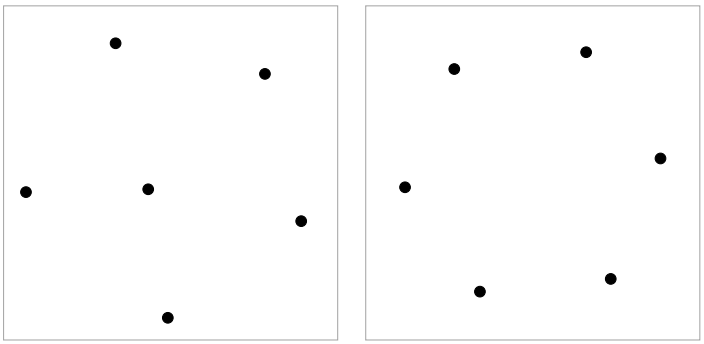


FIGURE 4.9 Increase of multi-information between $t = 0$ and $t = 5000$ for particle systems of different size ($m = 500$ samples, $r_c = 10$, $l = 1$ and $l = 3$ types using the same type specifications as in Figure 4.4 and Figure 4.6).

for certain numbers of particles. This has to do with the amount of similar geometric formations that are possible in a system consisting of particles of a single type, even though there is not an obvious systematic way in which this happens depending on the number of particles.

The small amount of organization for the six particle system, for example, can be explained by inspecting the individual samples at $t = 5000$. There are two different types of samples, forming either a pentagon, where the sixth particle is situated slightly off center, or a hexagon (see Figure 4.10). Now knowing the position of several particles still leaves a high entropy about the position of the remaining particles, as the configurations can be partially aligned, which means that there is a larger variation in shapes but, not an increase in variation in the individual particle positions.

There is another interesting effect in systems of particles of a single type. While it can be seen in Figure 4.9 that such a system organizes, both for a very small number of particles $n = 3$ and a larger collective of particles $n = 20$, this is not true if the cut-off radius r_c is decreased such that $r_c \leq 2r_{\alpha\alpha}$. So in the settings from above, but with a cut off radius of $r_c = 3$, the organization of the small system is still around 13 bit and the plot of the multi-information over time looks almost the same as for the same system with $r_c = 10$ (compare with Figure 4.4) while the organization of the large collective drops to around 2 bit (see Figure 4.11).



a) Pentagon configuration b) Hexagon configuration

FIGURE 4.10 Plot of the samples from the noise free three particle example at two different time steps. The particle configurations are shaded by sample. Therefore, it can be seen in a) that the outliers along the three axes belong to the same samples. This is a sign for correlation between the particles and hints towards a larger amount of multi-information.

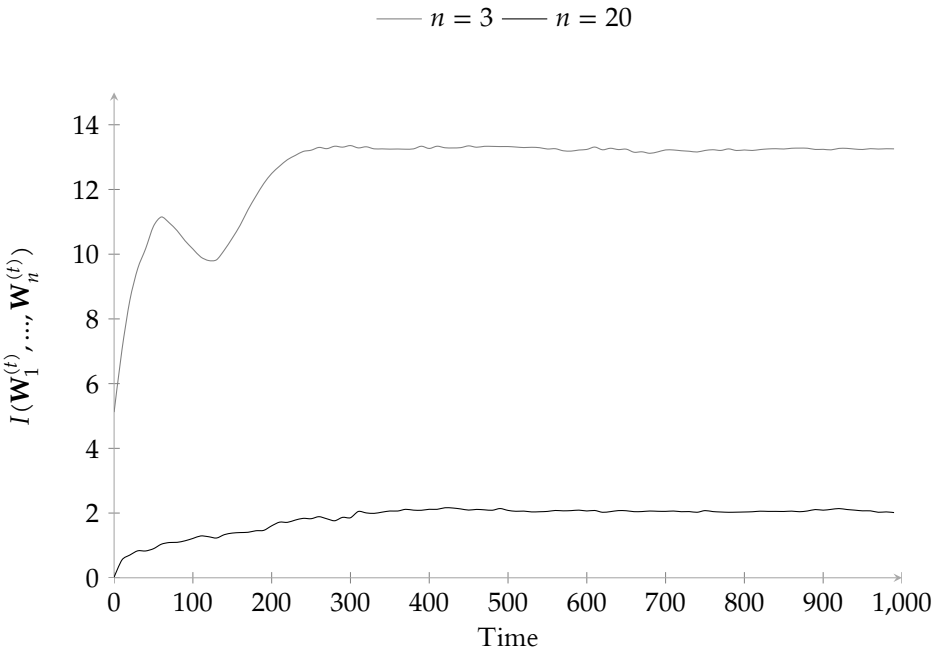


FIGURE 4.11 Multi-information between $n = 3$ and $n = 20$ particles of the same type with a smaller cut-off radius ($m = 1000$ samples, $r_c = 3$, $r_{\alpha\alpha} = 2.5$).

For the small system of three particles, reducing the cut-off radius does not change much (as long as the particles are initialized on a disc of radius $0.5r_c$, which they are in these experiments): In fact the equilibrium configuration is the same because every particle is still interacting with each other particle. This is not true for the larger collective of 20 particles. Here the resulting equilibrium configuration is always a regular grid and the self-organization is very low. This is due to two effects: The regular grid is also always roughly in the form of a disc, there is no variety in shapes, so the entropy for each particle

is already very low in general (after alignment), and more important, small perturbations in the grid structure are local and do not spread through the grid.

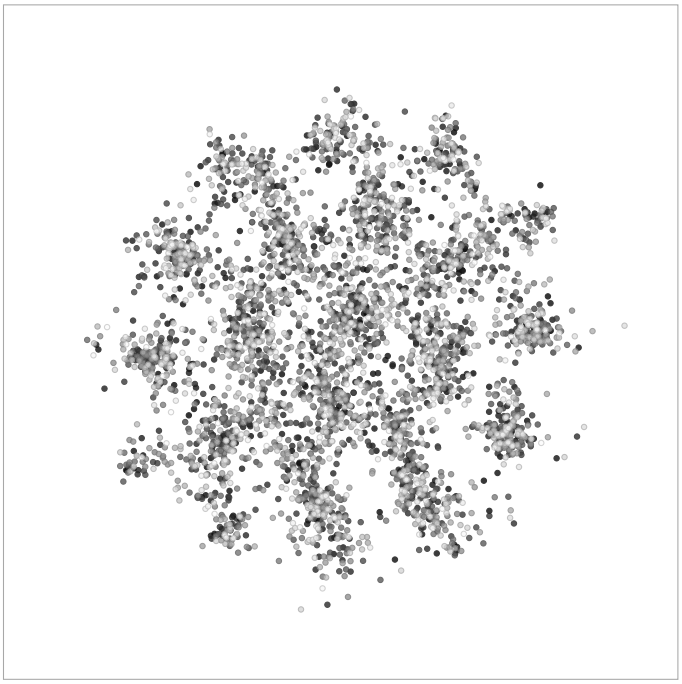
For the larger cut off radius ($r_c = 10$) the particles configure into two concentric regular polygons where the rotation of the inner polygon with regard to the outer polygon shows one degree of freedom (see Figure 4.12). For the small cut off radius noise and the initial random distribution of particles result in local variations, that are not correlated through the collective, thus increasing individual particles entropies but also the entropy of the whole system. Hence the lower value of multi-information in the equilibrium. For the larger cut off radius, noise is still reflected in local variations, but these are correlated through the collective resulting in a larger amount of organization (compare Figure 4.11 and Figure 4.9).

4.4.1 *Comparison of interaction types*

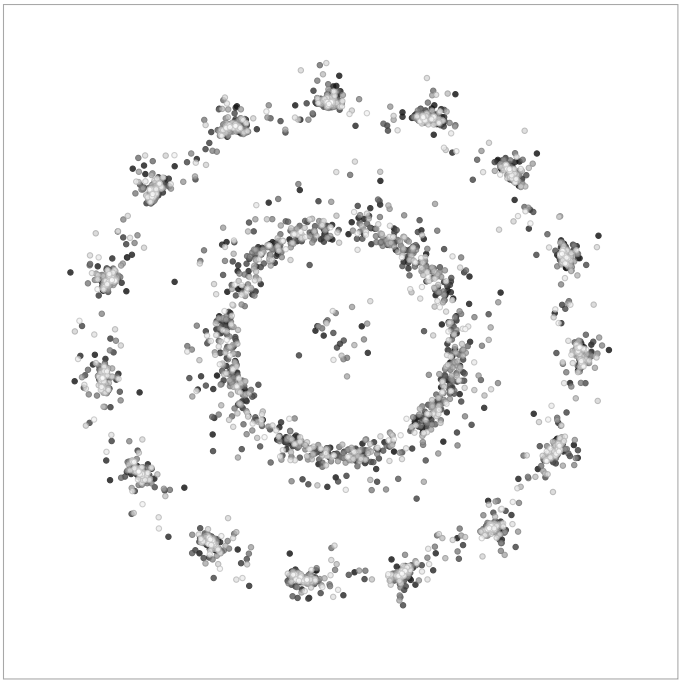
Simulations with $l = 3$ to 5 types and $n = 20$ to 75 particles almost always showed quantifiable self-organization reflected in an increase of multi-information (see Figure 4.8 for a typical example). It can be seen in Figure 4.13 that there is a decrease in self-organization with a larger number of types after an initial increase (for a fixed number of particles). This decrease can be attributed to the lower correlation in collectives with a high ratio of types to particles (this was already observed in the three particle example, compare Figure 4.6). A low ratio leads to clustering of particles of the same type, where the initial distribution of particles at $t = 0$ only shows a small influence on the final configuration. This is not the case if there are almost as many types as there are particles. In this case small fluctuations in the initial configuration can lead to quite different equilibrium configurations and thus a correlation amongst the particles in the equilibrium is harder to establish. These fluctuations are thus conserved over time. Nonetheless, in Figure 4.13 the maximal organization is reached by the systems consisting of four types which indicates that a certain amount of distinct types is actually helpful to reach a larger amount of organization.

Already in Figure 4.11 could be seen that the cut-off radius has a strong influence on the organization of a collective of a single type and reducing the radius below $2r_{\alpha\alpha}$ leads a large drop in organization. In the simulation above, the cut-off radius was set to $r_c = 7.5$ while the preferred mutual distances were $r_{\alpha\beta} \in [1.0, 5.0]$ and the initial particle configurations were drawn from a uniform distribution on a disc of radius 10. So initially not all particles are necessarily interacting with each other and the cut-off radius is smaller than twice the mutual preferred distance for some pairs of types.

In Figure 4.14 the increase of organization with an increase of r_c is shown. A decrease in organization for a high type to particle ratio, the effect that was mentioned above, can still be observed also for large cut-off radii. However, the decrease is more prominent for small values of r_c. Thus, long range interactions increase the organization of the particle system, but it seems that systems with only a few types can organize better when the dynamics dictate locally limited interactions.



a) $r_c = 3$



b) $r_c = 10$

FIGURE 4.12 Plot of all particles of all samples at time $t = 5000$, the system consists of 20 particles of a single type, shading of the particle denotes different samples. In a) a regular grid can be seen and in b) it can be seen that the outer ring has been much better aligned so that for each particle samples match more closely (denser clusters), while this is not possible for the inner ring of particles as their alignment related to the outer ring is a degree of freedom.

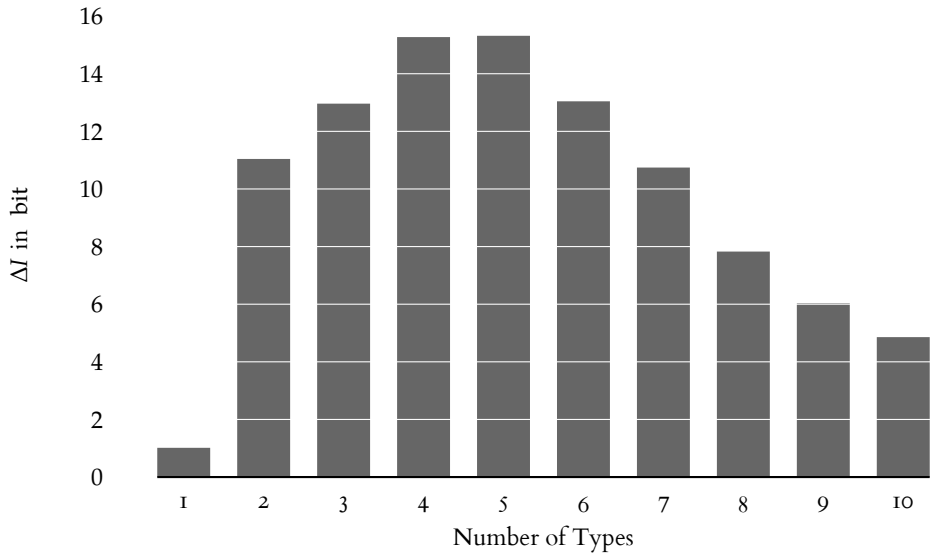


FIGURE 4.13 Increase of multi-information between $t = 0$ and $t = 1500$, for different numbers of types ($n = 20$ particles, $r_c = 7.5$ and $m = 250$ samples). Averaged over 30 randomly generated types with mutual preferred distance radii $r_{\alpha\beta} \in [1.0, 5.0]$ and $k_{\alpha\beta} \in [0.25, 0.75]$.

Spatial regularities are not a necessary condition for self-organization, but the mutual interactions define possible attractors to which the particles then organize. Because of the large number of different types compared to particles the structure is not (and cannot expected to be) regular. On the other hand, if the interactions are locally limited, for example because of a small cut-off radius, the self-organization is limited as well. Comparing this to the self-organization exhibited by systems with the same amount of particles, same local limitations on interactions, but considerably fewer different types, it possible to make the following observation: The increase of multi-information over time in these systems is much higher than in those being local and having as many types as particles.

To reach an increase in correlation (i.e. multi-information) among the particles, information needs to spread through the collective (Steudel and Ay, 2010). And hence, it is not surprising that long-range interactions lead to a lot of self-organization. What is, however, quite interesting is that self-organization is also possible in the case where the interactions are local but homogeneous. In these cases, where interactions are local, there are almost always smaller clusters interacting with each other. Each cluster shows a very regular structure and consists of particles of one type.

4.4.1.1 *Localization of organization*

If there is a cluster structure with spatially confined subsystems, there is a natural question: Is it possible to locate where the largest contribution to the organization is made? In Section 3.4, I explained that it is possible to decompose the multi-information of the observer variables into the several multi-information terms, that each measure the multi-information of a

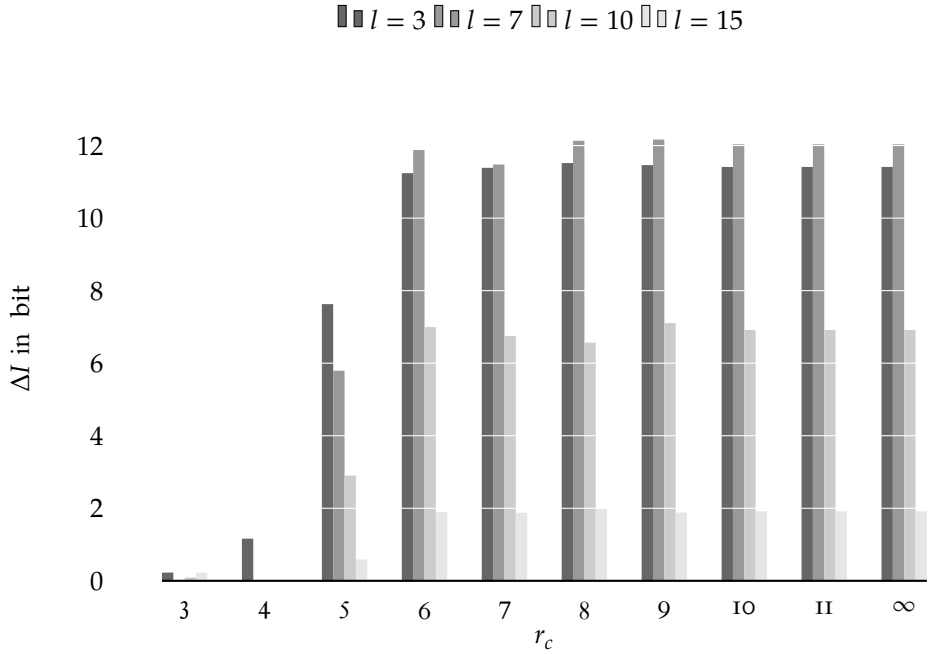


FIGURE 4.14 Increase of multi-information between $t = 0$ and $t = 1500$, for different cut of radii r_c and numbers of types l, ($n = 20$ particles and $m = 250$ samples). Averaged over 30 randomly generated types with mutual preferred distance radii $r_{\alpha\beta} \in [1.0, 5.0]$ and $k_{\alpha\beta} \in [0.25, 0.75]$.

subset of the observer variables, and one term that measures the multi-information between these coarse-grained joint observer variables. I now consider the joint random variable of all observers of a given type of particles as coarse-grained observers $\tilde{W}_1, ..., \tilde{W}_l$ (see Section 3.4.1), and calculate the multi-information individually. A general observation is that in every experiment it is possible to see organization on all levels. For a specific experiment with 5 different types of particles, I was able to observe the following: If the decomposition (coarse graining with respect to type) is normalized with respect to the multi-information for each time step, it can be seen that in the beginning of the experiment the relative contribution of each decomposition term still varies and it is possible to detect two different phases in the phase of organization from $t = 0$ to $t = 120$. In the first phase $t \in [0, 30]$ the largest contribution comes from interactions between type 0 particles. The next phase $t \in [30, 120]$ is dominated by a constant amount multi-information between the coarse grained observes, and an increase of type 2 observer multi-information (see Figure 4.15). This correlates with the dynamics where the interactions between particles of type 0 act with the largest forces, whereas the interactions between particles of type 2 are the smallest and thus, the organization of particles of type 0 will dominate the initial phase of organization. The coarsening with respect to type therefore allows to identify phases where different interactions are driving the organization process.

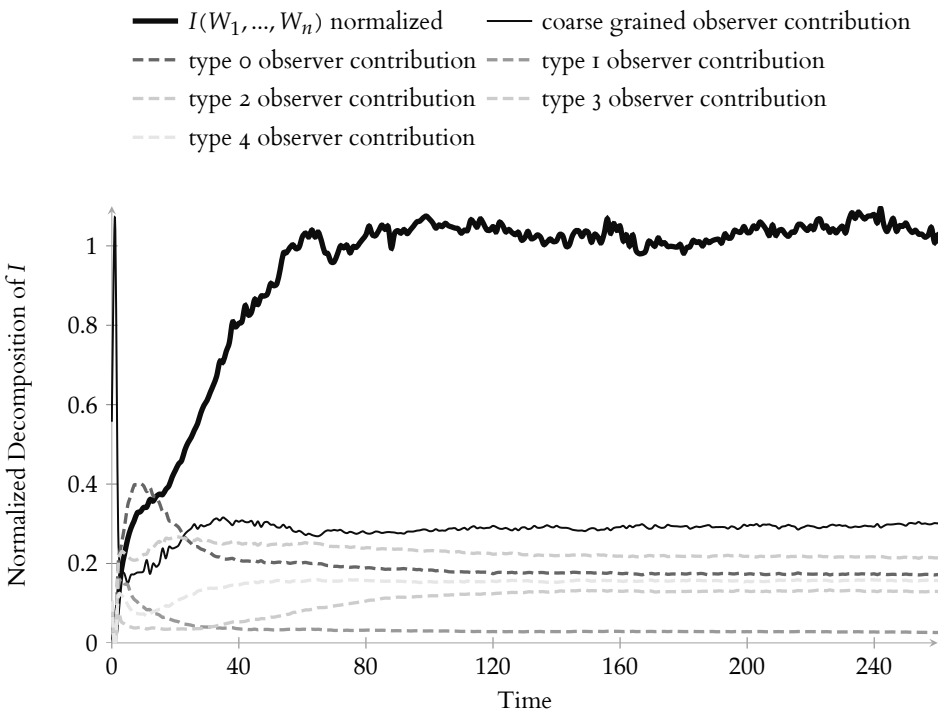


FIGURE 4.15 Contribution of the different terms of the decomposition normalized with the multi-information in each time-step. The total multi-information is normalized to fit the scale.

4.5 DISCUSSION

I used multi-information as a measure for self-organization and applied it to experiments of interacting particle systems. Estimations of multi-information were obtained using the Kraskov-Stögbauer-Grassberger estimator (Kraskov et al., 2004). As mentioned in Chapter 3, defining a measure for self-organization is not a straightforward task. With the definition used here one has to be careful in the choice of the observer variables. However, the results show that particle/type-based observers are a practicable approach to measure self-organization in spatial systems.

The first observation was, that a uniform collective (only one type) when forming regular grids only shows a small amount of measurable self-organization. If the cut-off radius limiting the interaction was increased, the collectives did not form regular grids anymore, but several concentric rings of particles with a rotational degree of freedom between them. The process from randomly distributed states to a regular grid structure is similar to the formation of crystals, which often is put forward as a classic example of self-organization. Even though, from a quantitative standpoint the self-organization of simple crystals seems to be not very high.

The main observation of the self-organization of particle systems concerned the variation of the cut-off radius r_c and the number of types in the particles. Here it is possible to see that given unconstrained interactions ($r_c = \infty$) the self-organization can be very high even if the

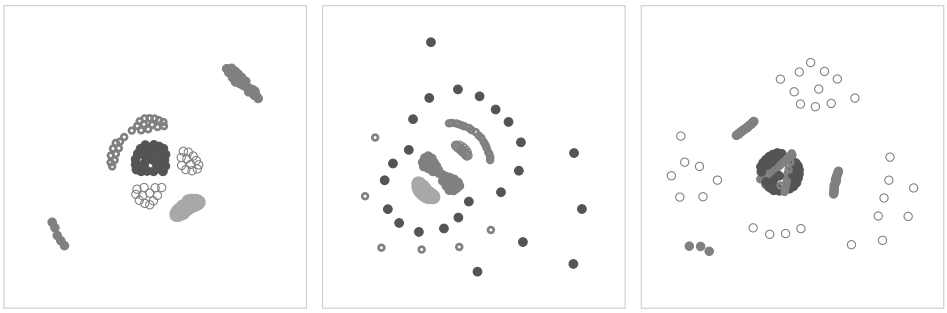


FIGURE 4.16 Examples of emergent structures in particle collectives.

particles have many distinct types with different mutual interactions. This was surprising insofar that the particle configuration in these settings do not show much spatial structure, and there is generally no emergent description in terms of clusters interacting with each other. However, the configurations show a lot of statistical structure, i.e. correlations, that the multi-information is able to detect. This can be related to the retrieval of spatial configuration of sensors using information-distance (Olsson et al., 2004). The distances are in this case represented by the $r_{\alpha\beta}$ radii and the experiment with $r_c = \infty$ is similar to the relaxation procedure that was used by Olsson et al. (2004) for the reconstruction of spatial structure. Another interesting point here is, that self-organization can occur without exhibiting a visually emergent spatial structure (e.g. 20 particles with 10 different types, large cut-off radius), this could support the idea put forward in (Shalizi, 2001) that self-organization and emergence are separate concepts.

Now, decreasing the cut-off radius r_c also decreases the observable self-organization (for a fixed number of types). This supports another assumption about self-organization: Information spread through the system is a crucial property of self-organizing systems. By limiting the cut-off radius, I am constraining the particles ability to transfer information through the system and therefore its ability to organize.

Now, if the number of types is decreased to three or four types (for a fixed value of r_c), the self-organization increases and one can observe emergent structures like balls enclosed in circles, layers of different types (see Figure 4.16). It seems that the emergence of clustered structures is a result of the way a system can achieve higher overall self-organization when interactions are locally constrained. Even with limited r_c, the homogeneity of the space as well as the homogeneity of local structures allow long-range structural interactions between groups of particles, which in turn allows to produce to a higher amount of self-organization of the whole system.

INFOGENESIS

5

» *All things physical are information-theoretic in origin and this is a participatory universe. Observer participancy gives rise to information; and information gives rise to physics.* «

JOHN ARCHIBALD WHEELER, Information, physics, quantum: The search for links

5.1 INTRODUCTION

The adaptation of an organism to its environment is driven by evolution, which takes place on a larger time-scale than the development of an individual organism, which is a guided process of self-organization towards an evolution-determined target form, obviously with some degrees of freedom for variations. The layout of body plans and cell differentiation is information that is implicitly encoded in the genetic code of every individual cell. This is different to what I have studied in Chapter 4, where the formation of shapes was induced by the dynamics of the environment, the particles were completely passive. This is not the case for living organisms where cells can sense molecules like neurotransmitters or hormone concentrations. More importantly, they also react to their sensor inputs by changing adhesion properties, cell motility, cell differentiation or even programmed cell death. The development of a living organism is, in a very abstract sense, a massive parallel information processing effort. And as Gregor et al. (2007) and Tkačik et al. (2009) show, apparently a very efficient one.

The terms self-organization or emerging structures are often considered to be opposite concepts to targets, optimization and control. I want to embrace a perspective where these concepts coexist. Self-organization, for example in a multi-agent system, does not contradict the existence of a (set of) target configuration(s), as long as the process towards the target configuration is autonomous. This makes it sensible that the target configuration is available to the agent collective from the beginning of the morphogenetic process, although possibly this information may be implicitly encoded in the policy of individual agents (as genetic code is in cells). If the target shapes are encoded in the physics that govern the collective as in Chapter 4 the agents can remain passive (i.e. they are just particles), but if the dynamics of the environment do not lead to an organization of the collective, the agents need to exert a certain level of control over their environment to reach the target configuration. This process can now be investigated using information-theoretic methods. I propose that, the transition from passive particles to reactive agents marks also the boundary between physics and biology and that the question of how dynamical systems can cross this boundary is possibly one of most prominent questions regarding the origin of life.

While nature provides evolution as an optimization process (roughly speaking and assuming a fixed environment), it is not always the ideal way to reach an optimum. If the interest lies only in the result and not in a study about evolutionary processes, other optimization methods can be used to obtain information-theoretic limits on control, self-organization and coordination. One might ask at this point whether such an optimized system has

anything to do with self-organization. I propose to answer this question positively. The optimization process is only used to obtain a collective that is optimally adapted, i.e. reaches a target configuration in an optimal way (given some constraints). The optimization of collective behaviour (the policy) is offline in the sense that the agent is not optimizing its own policy by exploring the world and thus learning. The collective equipped with such an optimal policy, however still organizes autonomously. Similar as evolution is an offline optimization guiding the development of living organisms, albeit this does not imply that evolution is actively controlling the development of an individual organism.

To begin, I will revisit the concept of relevant information as introduced by Polani et al. (2006) and introduce the work on information-theoretic control theory by Touchette and Lloyd (2004,2000). I will then discuss the embodiment of agents and the representation of agent collectives in the perception-action loop, as this will be required to create an information-theoretic perspective on morphogenesis.

5.2 INFORMATION-THEORETIC CONTROL THEORY

Touchette and Lloyd (2004) were the first to establish a formal link between control theory and information theory. Control theory is concerned with controllable dynamical systems, i.e. systems with an input. Depending on whether the state of the system is fed back into the controller or not, the control is called closed-loop or open-loop control. CBNs are ideal to formalize stochastic control systems (see Figure 5.1). Here, W denotes the system state, C the controller and W' the state of the system after control was applied. The random variables of CBNs in this chapter are assumed to be defined on finite spaces. A first application of information theory is a formal definition of open-loop and closed-loop control via mutual information: If $I(W;C) = 0$ the control is called open-loop, and hence the arrow in the CBN can be omitted, if $I(W;C) > 0$, it is called closed-loop.

Touchette and Lloyd (2004) relate control theoretic concepts like controllability and observability to information-theoretic formulations. Here, I will be mainly concerned with their work on optimal control as this gives some valuable insight into the information processing of control systems. The entropy reduction for an open-loop control system is defined by them for each control state $c \in C$ as follows:

$$\Delta H^{c}_{\text{open}} := H(W) - H(W'|C = c) \tag{5.1}$$

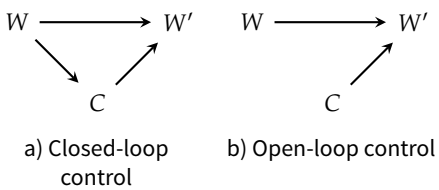


FIGURE 5.1 CBN of control systems, W is the random variable that denotes the system state, C the controller and W' the system after control was applied.

where $H(W'|C = c)$ is the entropy of W' controlled by an open loop controller C. The maximal entropy reduction achievable for an open-loop controller is now given by

$$\Delta H^{\max}_{\text{open}} = \max_{p(w)\in\Delta(W), c\in\mathbb{C}} \Delta H^{c}_{\text{open}}. \tag{5.2}$$

In (Touchette and Lloyd, 2004) it is shown that any non-deterministic open-loop controller can only achieve as much entropy reduction as a deterministic one, therefore $\Delta H^{\max}_{\text{open}}$ is the maximum achievable entropy reduction, for any distribution of W given the dynamics $p(w'|w, c)$.

In a similar way the actual entropy reduction of a closed-loop control is defined as

$$\Delta H_{\text{closed}} = H(W) - H(W'). \tag{5.3}$$

One of the main results of (Touchette and Lloyd, 2004) is the following inequality

$$\Delta H_{\text{closed}} \leq I(W; C) + \Delta H^{\max}_{\text{open}}. \tag{5.4}$$

This is a very elegant result, because it states that for every additional bit of entropy reduction in the system (compared to the best possible open-loop controller), the controller needs to take one bit from the system and process it. A related observation was made by Klyubin et al. (2004) who showed that information about the initial state of an gradient following agent passes through the agents action sequence.

Following (Touchette and Lloyd, 2004), a closed-loop controller is considered optimal or maximally efficient if

$$\Delta H_{\text{closed}} - \Delta H_{\text{open}} = I(W; C) \tag{5.5}$$

where ΔH_{open} is the actual entropy reduction, i.e $H(W) - H(W')$ but with $p(w', w, c) = p(w'|w, c)p(w)p(c)$, where $p(c)$ is the marginal of C, hence making W and C independent in the transition of the system.

Now, it is in theory possible to maximize the entropy reduction for closed-loop control to maximize control over the system. At the same time there are systems where information processing is costly, so it is natural to assign a cost on information processing. I propose the following trade-off formulation of the entropy reduction, in the spirit of a rate-distortion minimization (Cover and Thomas, 2006),

$$\min_{p(c|w)} I(W; C) - \beta \Delta H_{\text{closed}}. \tag{5.6}$$

For $\beta \to 0$ the resulting controller will approximate the best possible open-loop controller because $I(W; C) \to 0$ whereas for $\beta \to \infty$ information processing is only of secondary interest and entropy reduction will be maximal. I did not prove whether the resulting controllers are also maximally efficient in general or not, as this is not important in what

follows. If $\Delta H^{\max}_{\text{open}} = 0$, that is open-loop control is ineffective, and maximally efficient controllers exist for all possible values of $I(W;C)$, then the solution of Eq. (5.6) for any β is obviously from this set of maximally efficient controllers.

In Eq. (5.6) the reduction of entropy was the only goal (with an information processing constraint). In practice this would be a rather unspecific goal, although in conjunction with the oberserver self-organization introduced in Chapter 3, this might be used to maximize self-organization in collectives, albeit in a unguided way. Most control theoretic algorithms optimize control towards a desired system state. A subfield within the field of optimal control developed into the field of dynamic programming and reinforcement learning Bellman and Kalaba (1965) and Sutton and Barto (1998), where control is optimized to maximize an arbitrary reward function for state transitions. Now, a small change in Eq. (5.6) relates reinforcement learning and information-theoretic control. This was initially proposed by Polani et al. (2006) under the label of relevant information which will be introduced in the next section. Related problem are considered in (Saerens et al., 2009 and Todorov, 2009). In the former optimization is concerned with randomized shortest paths in networks with information constraints, thus there is no model of embodied agents, whereas in the latter the information-theoretic constraints are imposed on the general state transition and not on the information processing of the controller. All these problems are similar to rate-distortion problems (Cover and Thomas, 2006), however with a distortion function that depends on the conditional distribution of the channel, which needs special attention for solving these problems, as will be seen later in this chapter.

Hence, although Eq. (5.6) has not much practical relevance because blind entropy reduction is rarely desired, it is true that most controllers reduce entropy and thus Eq. (5.4) is an important limit for control systems. Furthermore, as I will show in the next sections, it is possible to consider embodied agents as controller of their environment and hence use the information-theoretic limits of control systems to obtain limits for the information processing of embodied agents.

5.3 RELEVANT INFORMATION

Instead of the single time-step control system, I will now consider the perception-action loop as introduced in Section 2.1.10. The perception-action loop in its most simple form is a closed-loop control system unrolled over time (see Figure 5.2). In contrast to the perception-action loop as it was introduced before, there are no random variables for the sensors and it is assumed that the agent has access to the full world state W_t at every time-step t. The controller is now the agent's actuator and hence denoted A_t. The random variables of the perception-action loop are indexed by time and the distribution P_{W_t} denotes the actual distribution at time t and W_t as well as A_t are defined on the world state space $\mathcal{W}$ and action space $\mathcal{A}$ respectively. The random variables W, W' and A (also defined on $\mathcal{W}$ and $\mathcal{A}$ respectively) stand for a typical transition in the perception-action loop where the

distribution P_W is a time average and $P_{W'|A,W}$ is equal to $P_{W_{t+1}|A_t,W_t}$ as the world state transition are time translation invariant (the world has constant dynamics). For a stationary world this does not change much, as $P_W = P_{W_t}$ for all t. If stationarity is not given, a naive approach would be to assume P_W to be uniform. This can always be done if the calculation of a time average world state distribution is not possible, however the results might be skewed heavily if the uniform prior is used. In episodic scenarios it is possible to average episodes over time. The actual calculation will be shown in Section 5.7.

Now consider a utility function $U^\pi(w,a)$, that assigns a value to each state action pair (w,a) if the agent acts according to some policy π. In traditional reinforcement learning the agent's goal is now to maximize the expected utility, by changing its own policy π (Sutton and Barto, 1998), which is used as a shorthand notation for conditional distribution $P_{A|W}$.

$$\max_{P_{A|W}} \mathbb{E}[U^\pi(w,a)]. \tag{5.7}$$

where the expectation is averaging over $p(w,a)$.

However, this maximization completely ignores the information processing burden the agent has to carry. The resulting policy could be overly complex as long as it is optimal with respect to utility. To address this problem Polani et al. (2006) introduced the following optimization term:

$$\min_{P_{A|W}} I(W;A) - \beta \mathbb{E}[U^\pi(w,a)]. \tag{5.8}$$

The result is a minimization of mutual information between world state and agent actuator with the expected utility as a constraint. Here, one needs to be aware that, unless W is assumed to be uniformly distributed, a change of the agent's policy $P_{A|W}$ also changes the distribution of P_W as it is a time average.

By varying β a trade-off between information processing and performance (expected utility) is made. For $\beta \to 0$ the policy approximates, similar as above, the optimal open-loop control policy, whereas $\beta \to \infty$ converges to an optimal policy that requires the least amount of information (per time step) from the world state. Plotting $I(W;A)$ against $\mathbb{E}[U^\pi(w,a)]$ results in trade-off curves as illustrated in Figure 2.3.

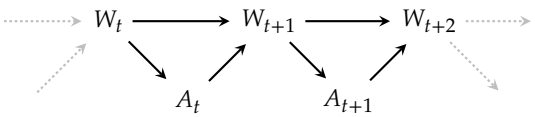
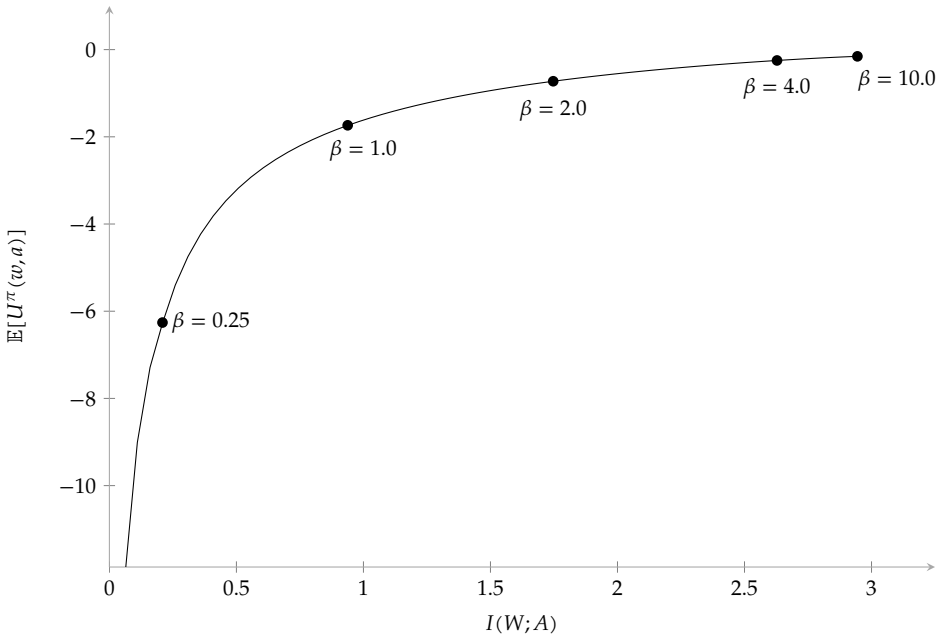
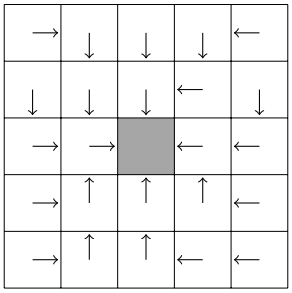


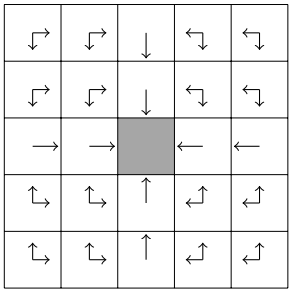
FIGURE 5.2 Illustration of the CBN of the perception-action loop of a memoryless agent with full access to the world state.



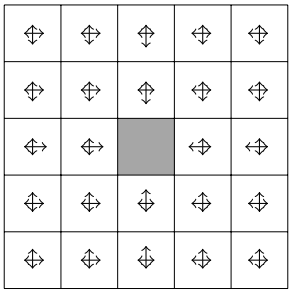
a) Information-utility trade-off curve for a simple goal finding task in a grid world.



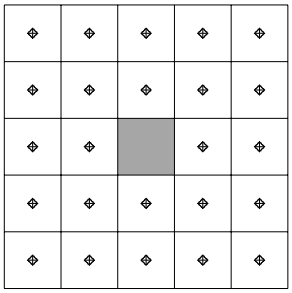
b) Optimal policy, without information constraint



c) Relevant information policy at $\beta = 10$



d) Relevant information policy at $\beta = 0.25$



e) Relevant information policy at $\beta \to 0$

FIGURE 5.3 Illustration of the relevant information formalism for a simple goal finding task in a 5×5 grid world. In a) the trade-off curve between relevant information and performance is shown, b) shows an optimal policy that can be the result of an optimization without any information constraint and c) - d) show policies for different value of β.

5.3.1 *Value and Utility*

The utility function encodes a reward structure that defines the actual goal(s) of the agent. For every step the agent gets a reward that is determined by a reward function $r(w', a, w)$ which depends on the current state, the action taken and the state of the world after the action was executed. The reward function and the perception-action loop now define a Markov Decision Process (MDP). The utility function $U^\pi(w, a)$ is then defined recursively via a state value function $V^\pi(w)$ that gives the expected future reward while currently being in some state $w \in \mathcal{W}$ and following the constant policy π ($=P_{A|W}$):

$$V^\pi(w) = \sum_a p(a|w) U^\pi(w, a), \tag{5.9}$$

$$U^\pi(w, a) = \sum_{w'} p(w'|a, w) \left(r(w', a, w) + V^\pi(w') \right). \tag{5.10}$$

5.3.2 *Blahut-Arimoto Iteration*

The definition of the state value function is recursive and the correct value function is a fixed point of this equation. Classic reinforcement learning now states that iterating the recursive definition of the value function converges to the correct value function for a given policy (Sutton and Barto, 1998). It is now possible to combine the value iteration with the Blahut-Arimoto algorithm (Blahut, 1972), which on itself can be used to solve rate-distortion problems (Cover and Thomas, 2006). This combination was, to my knowledge, first introduced by Polani et al. (2006). The Blahut-Arimoto iteration is given by

$$p_{k+1}(a|w) = \frac{p_k(a)}{Z_k(w, \beta)} \exp(\beta U^\pi(w, a)), \tag{5.11}$$

$$p_{k+1}(a) = \sum_w p_k(w) p_k(a|w), \tag{5.12}$$

where k denotes the iteration step, $Z_k(w, \beta)$ is a normalisation term and $\beta > 0$ the trade-of between optimality and relevant information as introduced above. Now the iteration is alternated with an update of the state probabilities $p_k(w)$ according to the current policy and a value iteration to get a consistent utility U^{π_k}. For the combined policy iteration the iterations steps are then done in the following order

$$p_k(a|w) \to p_k(w) \to V^{\pi_k} \to U^{\pi_k} \to p_{k+1}(a|w). \tag{5.13}$$

The algorithm then minimises the term in Eq. (5.6) for a given value of β and returns an optimal policy for the agent given the constraints and thus allows to calculate the relevant information for different information processing capabilities.

5.4 EMBODIMENT & PERCEPTION-ACTION LOOPS

The perception-action loop was introduced in Section 2.1.10 as a CBN capturing the embodiment of an agent (see Figure 5.4). In the case of the simplified perception-action

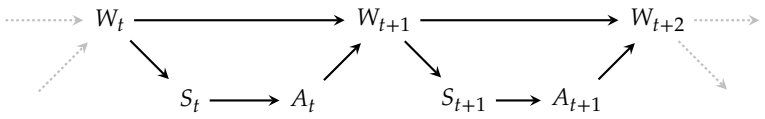


FIGURE 5.4 Illustration of the CBN of the perception-action loop of a memoryless agent.

loop with full world access (see Figure 5.2), this is only partially true. For example, a policy $P_{A|W}$ obtained from the relevant information optimization also determines how the agent accesses information from the environment and this is implicitly describing its sensors, the sense that an information bottleneck (Tishby et al., 1999) could be used to obtain a distinct sensor variable S based on the policy $P_{A|W}$ (maximizing the information S contains about A and minimizing the information S contains about W).

Besides sensors, many organisms have an internal information storage and information processing apparatus (of which the brain is an important part) which I call memory. In the perception-action loop it is denoted by the random variable M_t. It is separated from the dynamics of the external world via sensors and actuators. While in Figure 5.4, sensor and actuator are Markovian with respect to the world state, this is not the case for memory which operates parallel to the world. The perception action-loop for an agent with memory (see Figure 5.5) is actually symmetric along the sensor-actuator axis from an information-theoretic point of view. In the reinforcement learning perspective from above, the perception-action loop with memory, together with a reward function define a partially observable Markov decision process (POMDP) with finite discrete belief states (Hansen, 2008).

The answer to the question of how to detect embodiment within a dynamical system is far from simple and part of ongoing research (Biel and Polani, 2012). The importance of embodiment for the cognitive process has been emphasized in Artificial Intelligence as well as philosophy, though there is no exact definition of what embodiment is (Clark, 1998, Pfeifer and Scheier, 2001, Gallagher, 2005 and Pfeifer and Bongard, 2007). In the context of the perception-action loop an embodied agent represents an entity that performs information processing parallel to everything else that goes on in the world, while interacting with the world through some prescribed communication channels. Depending on the actual model of the perception-action loop, there are several implicit assumptions that are usually taken for granted, especially in spatial environments.

Consistency of Actions: For example in a two dimensional environment, the actions are usually modelled in a consistent way, such that an action labelled 'left' actually moves the agent to the left as long as no obstacle is encountered, but never to the right, upwards or downwards. In this case this means that actions are translation invariant. This consistency has implications on the information processing as Polani (2011) showed. The reader familiar with differential geometry might be reminded of an affine connection by this assumption, which allows parallel transport along a curve on a Riemannian manifold (Lee, 1997). The

idea, that an agent can choose the same actions in every world state, seems similar to the isomorphism of tangent spaces that follows from an affine connection, even though the manifold itself can look different locally, similar like the reaction of the environment to actions of the agent can depend on the state of the world. But an affine connection not only identifies the tangent spaces with each other, furthermore the exponential map, connects the tangent vector to the manifold, so, even with local curvature, the tangents are consistently identified along curves on the manifold. Consistency thus means, in a spatial system, that actuators can be considered to be global and agents are able to select actions independent of the state of the world, while their effect is local but consistent (it depends on the current state of the world as given by $P_{W_{t+1}|A_t,W_t}$).

Locality of Sensors: Sensors are similar, usually their cause is local: their state depends on the state of the world and in many models the sensor is simply a local read-out of some part of the world state (i.e. a projection). Nonetheless sensors also need to show some consistency, which introduces a global connection between sensor variables. The lack of consistent sensors would possibly influence information processing of the agent in a similar way as inconsistent actuators do in (Polani, 2011).

Conditional Stationarity: While this assumption is often not true for biological systems, many models assume that the conditional distributions in a perception-action loop are time translation invariant. This has two implications: The embodiment of the agent is constant and it does not perform any learning. So, while conditional stationarity does not hold for more complex organisms, it suffices as an assumption on a lower level where only a minimal amount of cognition is involved and changes in the embodiment are considered slow in comparison with more complex organisms. For the model of monocellular organisms for example, the perception-action loop of an agent with memory as illustrated in Figure 5.4 can be a good fit with the exception of cell-division which however has not yet been incorporated into the perception-action loop framework.

The first two consistency assumptions are rather vague and are currently missing a formal framework. I believe that there are many analogies to differential geometry (Lee, 1997) and possibly discrete differential geometry (Grinspun et al., 2006) provides helpful insights towards a formal definition of a consistent embodiment within the perception-action loop. Here, it suffices to note that at this point, that a physical embodiment, as it is often considered, is not entirely informational. There are some natural assumptions underlying the sensor and actuator random variables that are physical. While a purely information-theoretic embodiment, which is implicitly encoded in the transition $P_{W_{t+1}|A_t,W_t}$, could in theory look quite unlike any biological organism.

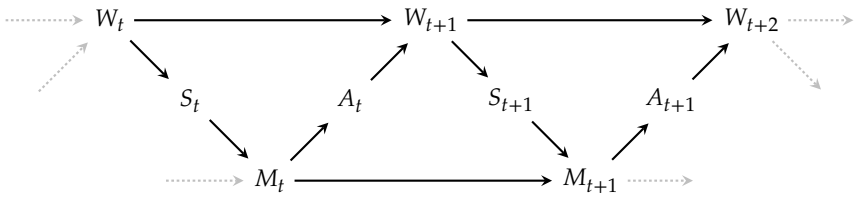
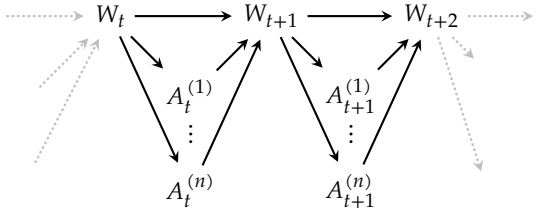
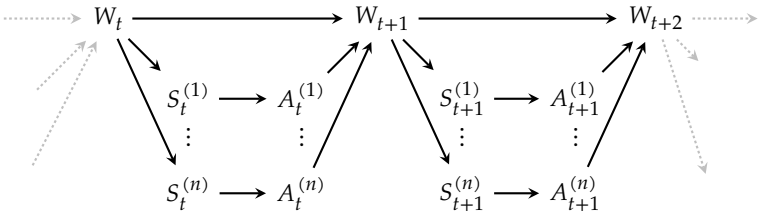


FIGURE 5.5 Illustration of the CBN of the perception-action loop of an agent with memory.



a) Agents with full world access



a) Agents with sensors

FIGURE 5.6 Illustration of the CBNs of the perception-action loops of a collective of n memoryless agents.

5.5 MULTI-AGENT RELEVANT INFORMATION

The perception-action loop formalism easily extends to multi-agent systems. Figure 5.6 shows the extensions of the memory-less perception-action loops to a multi-agent system of n agents. The sensor and actuator random variables are now denoted $S_t^{(1)}, ...S_t^{(n)}$ and $A_t^{(1)}, ...A_t^{(n)}$ respectively. Furthermore, in the multi-agent setting, let $S_t = (S_t^{(1)}, ...S_t^{(n)})$, with $s_t = (s_t^{(1)}, ..., s^{(n)})$ and $A_t = (A_t^{(1)}, ...A_t^{(n)})$ with $a_t = (a_t^{(1)}, ..., a^{(n)})$ respectively. In practice the sensors and actuators of agents in a collective are often identical, however this is not a requirement of the formalism. As above, the time averaged random variables in the perception-action loop are denoted by the index-less variables.

Now it is possible to use the relevant information formalism to gain informationally optimal policies for a given reward function $r(w', a, w)$. The reward is depending on the current state of the world, the next state of the world and the joint action $a = (a^{(1)}, ..., a^{(n)})$ of all agents. The relevant information minimization term for an agent collective with full-world access now looks almost the same except for a minor change. As the agents are acting individually, there is an additional constraint that the joint policy needs to fulfil:

$$p(a|w) = p(a^{(1)}|w) \cdots p(a^{(n)}|w), \tag{5.14}$$

namely, given the world state all agent policies need to be independent of each other. Then the relevant information minimization is as follows

$$\min_{P_{A|W}, p(a|w)=p(a^{(1)}|w)\cdots p(a^{(n)}|w)} \left(I(W;A) - \beta\mathbb{E}[U^{\pi}(w,a)]\right). \tag{5.15}$$

In case of an agent collective that does not have full world access but where each agent is equipped with a sensor the minimization term changes. The information processing of each agent happens between sensor and actuator and therefore, the sum of all mutual informations is minimized. At the same time, the utility is still defined via the state of the world and the joint action:

$$\min_{P_{A^{(1)}|S^{(1)}}, \ldots, P_{A^{(n)}|S^{(n)}}} \left(\sum_{i=1}^{n} I(S^{(i)};A^{(i)}) - \beta\mathbb{E}[U^{\pi}(w,a)]\right). \tag{5.16}$$

The collective joint policy given the world state is now

$$p(a|w) = \left(\sum_{s^{(1)}} p(a^{(1)}|s^{(1)})p(s^{(1)}|w)\right) \cdots \left(\sum_{s^{(n)}} p(a^{(n)}|s^{(n)})p(s^{(n)}|w)\right). \tag{5.17}$$

Now the iteration that was used for a single agent with full world access to compute relevant information can also be used to compute relevant information in multi-agent scenarios where the agents have full world access. The iteration is now alternated between the agents. For each agent a value iteration and a Blahut-Arimoto iteration is performed using the current policy of the other agents as a predictor in the utility update. That is for the value iteration of each agent, the policy of the collective is assumed to be independent $P_{A^{(1)}|S^{(1)}}, \ldots, P_{A^{(n)}|S^{(n)}}$. This means each agent can anticipate the action of the other agents as it is intrinsically aware of their policies. This makes sense if the agents are assumed to be identical and have a shared evolutionary history. In other cases it is possible in the value iteration step to replace the policies of all other agents by uniform distributions and assume no prior knowledge of the other agents' actions. The general scheme of iterations is now (similar to the iterations of the multivariate information bottleneck (Friedman et al., 2006))

$$\begin{aligned} p_k(a^{(1)}|w), \ldots, p_k(a^{(n)}|w) \to p_k(w) \to V^{\pi_k^1} \to U^{\pi_k^1} \to p_{k+1}(a^{(1)}|w) \to \ldots \\ \ldots \to V^{\pi_k^n} \to U^{\pi_k^n} \to p_{k+1}(a^{(n)}|w). \end{aligned}$$

First, there are the policies for each agent from which the common environmental state distribution is calculated. This is followed by a value iteration step for the first policy and a Blahut-Arimoto update that gives the new policy for the first agent. Using this policy and the current policy of all other agents as a predictor, the value iteration step for the next agent is done, again followed by a Blahut-Arimoto step, and so on until the last agent is reached at which point an iteration ends. This iteration now converges to optimal policies for all agents while minimizing $I(W;A^{(i)})$ of each agent. As the value iterations are independent of each other for the collective (in so far as each agent has its own value and utility function), the reward is scaled by the factor $\frac{1}{n}$, so that the total reward for

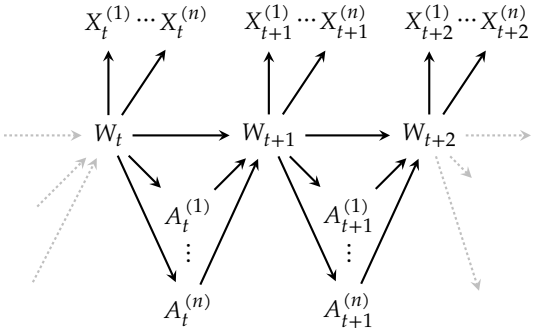


FIGURE 5.7 Illustration of the CBNs of the perception-action loops of a collective of n memoryless agents.

the collective per step is still $r(w', a, w)$, where $a = (a^{(1)}, ..., a^{(n)})$ as mentioned earlier. A simple calculation shows that this also means that the performance of the policy of the whole collective $\left(p(a|w) = p(a^{(1)}|w) \cdots p(a^{(n)}|w)\right)$ is equal to the sum of the individual performances in the collective

$$\mathbb{E}[U^{\pi}(W, A)] = \sum_{i=1}^{n} \mathbb{E}[U^{\pi^i}(W, A^{(i)})]. \tag{5.18}$$

This will be important later when the performance of an agent collective with independent agents $\left(p(a|w) = p(a^{(1)}|w) \cdots p(a^{(n)}|w)\right)$ is compared to a collectives with shared control of the agents, that means the collective is seen a a single distributed agent with a policy $P_{A|W}$.

5.5.1 *Observable Agent Collectives*

To connect the measure of O-self-organization with the relevant information formalism and the perception-action loop, each agent needs to have an observer variable. For this each agent is observed by a variable $X_t^{(i)}$ where $\mathcal{X}_1 = ... = \mathcal{X}_n = \mathcal{X}$ with $\mathcal{X}$ being the state space of an individual agent (e.g. position, type). This extends the multi-agent perception-action loop as illustrated for the agent collective with full world access in Figure 5.7

The state representing location and other agent related states (for example the type of an agent as in Chapter 4), will now be changed by the agent as part of an actuation. If agents can sense and act on their own state and there is no interaction between state and the rest of the world, the state acts as an internal memory and a different perception-action loop model makes more sense (see Section 5.4). If however, the world dynamics interact with the state or other agents can sense an agent's state like their position, the state is more of an external memory and is part of the world state.

A requirement for the agent observers is that they are fully determined by the world state, that is

$$H(X_t^{(1)}, ..., X_t^{(n)}|W_t) = 0 \text{ for all } t, \tag{5.19}$$

as well as being time translation invariant, i.e. $p(x_t^{(1)}|w_t) = p(x_{t+1}^{(1)}|w_{t+1})$ if $w_t = w_{t+1}$ for all i and all t. This means that for each agent there is a projection function $\rho_i : \mathcal{W} \to \mathcal{X}_i$ extracting the agent's state from the world state.

If the converse is also true, namely

$$H(W_t|X_t^{(1)}, ..., X_t^{(n)}) = 0 \tag{5.20}$$

the collective will be called isolated. If the collective is not isolated it is helpful to model the world as a joint variable of agent state as well as a random variable R_t capturing the rest of the world, that is $W_t = (X_t^{(1)}, ..., X_t^{(n)}, R_t)$.

5.6 RELEVANT INFORMATION & SELF-ORGANIZATION

In the case of an isolated agent collective the observer variables can be used to measure the self-organization of the whole system via observers as introduced in Chapter 3. If the agent collective is not isolated, it is still possible to measure the organization of the collective itself, but possible correlations to other parts of the world need to be considered, for example by having a random variable encoding the state of the rest of the world as mentioned above. In what follows I will consider an isolated system for simplicity, though the general ideas should be easy to transfer to models where the world state consists of more than the agents' locations and states.

Relevant Information provides a lower bound on the information that an agent collective at least needs to process on average per time step to reach a configuration at a given performance level. Now I want to look into the amount of information a collective needs to process to reach a specific amount of self-organization. Here, I will assume, that for the system, for each transition W_t to W_{t+1},

$$\Delta H_{\text{open}}^{\max} = 0, \tag{5.21}$$

that is, there is no joint action of the collective that independently of the state of the system decreases its entropy. This is unlike to the particle systems in Chapter 4 where the dynamics of the environment already did reduce the entropy of the whole system. On the other hand it also means that there is no noise in the system or more correctly that entropy production (noise) and reduction (e.g. particle dynamics) are balanced unless the collective performs information processing.

Now there are two drives, that can increase organization, the increase of the individual entropies $H(X_t^{(i)})$ and the decrease of the joint entropy $H(X_t^{(1)}, ..., X_t^{(n)})$. The increase of individual entropies, while keeping the joint entropy constant, requires something like correlated entropy production, this would require highly non-local world dynamics or a high degree of coordination, i.e. agents producing correlated entropy, for example by performing correlated random walks. Coordination and shared control will be the topic of

the next section. In many scenarios the individual entropies are actually decreasing over time, because the collective was initialized with maximal or high individual entropies. With the assumption that $H(X_t^{(i)}) \geq H(X_{t+1}^{(i)})$ it follows from (5.4) that

$$I(X_{t+1}^{(1)}, ..., X_{t+1}^{(n)}) - I(X_t^{(1)}, ..., X_t^{(n)}) \leq \Delta H_{\text{closed}}^t \leq I(W_t; A_t). \tag{5.22}$$

where $\Delta H_{\text{closed}}^t = H(X_t^{(1)}, ..., X_t^{(n)}) - H(X_{t+1}^{(1)}, ..., X_{t+1}^{(n)})$ as introduced above. With the assumption that $I(X_0^{(1)}, ..., X_0^{(n)}) = 0$ it follows that

$$I(X_t^{(1)}, ..., X_t^{(n)}) \leq \sum_{\tau=0}^{t-1} I(W_\tau; A_\tau). \tag{5.23}$$

This result is relevant, as it limits the amount of self organization, that is achievable in t time steps by the total information processed in this time. For systems, where $\Delta H_{\text{open}}^{\max} = 0$ does not hold, this generalizes to

$$I(X_t^{(1)}, ..., X_t^{(n)}) \leq t\Delta H_{\text{open}}^{\max} + \sum_{\tau=0}^{t-1} I(W_\tau; A_\tau). \tag{5.24}$$

Furthermore, if the agents do not have full world access, it follows from the data processing inequality (Cover and Thomas, 2006) that also

$$I(X_t^{(1)}, ..., X_t^{(n)}) \leq t\Delta H_{\text{open}}^{\max} + \sum_{\tau=0}^{t-1} \sum_{i=1}^{n} I(S_\tau^{(i)}; A_\tau^{(i)}). \tag{5.25}$$

This also means that any organization exceeding the level induced by the dynamic of the environment (accounted for by the term $t\Delta H_{\text{open}}^{\max}$) needs to be processed by the collective.

5.7 EPISODIC TASKS & SHAPES AS GOALS

For a morphogenetic process of guided self-organization, the guiding consists of a certain target configuration to which the collective should organize itself. This can simply be a shape, i.e. an n-tuple of coordinates, but also include the state of agents. For example if agents are of three different types, as in the classic French-flag example (Wolpert, 1969), would lead to a target where the agents form a rectangle with three stripes, where for each stripe all the agents in it have to have the same type. A target configuration is therefore simply an element $(x_1, ..., x_n) \in \mathcal{X}_1 \times ... \times \mathcal{X}_n$.

Now it is possible to have more than one target configuration. For example if small variations are unimportant, several configurations could be considered as targets. The set of target configurations will be denoted $\tilde{\mathcal{X}}$. Moreover, as in Section 4.1.2, there are transformations of $\mathcal{X}_1 \times ... \times \mathcal{X}_n$ which leave the configuration invariant with respect to the specification of a target. The set of configuration invariant transformations, denoted F, depends, except for the identity map, on the actual model of the world, but in most cases it will include translations, rotations and permutation of agents that share the same state.

Equipped with the set of target configurations $\tilde{\mathcal{X}}$ and the set of invariant transformations F it is now possible to define a reward function $r_{\tilde{\mathcal{X}};F}$ on $\mathcal{W} \times \mathcal{A} \times \mathcal{W}$. The reward is negative, whenever the collective is not reaching one of the target configurations and zero when a target is reached:

$$r_{\tilde{\mathcal{X}};F}(w', a, w) = -\delta(w') \tag{5.26}$$

where

$$\delta(w) = \begin{cases} 0, & \text{if } \exists (x_1, ..., x_n) \in \tilde{\mathcal{X}} \wedge f \in F : \rho(w) = f(x_1, ..., x_n) \\ 1, & \text{else} \end{cases}$$

with $\rho(w_{t+1}) = (\rho_1(w_{t+1}), ..., \rho_n(w_{t+1}))$ being a tuple of projections to the observer variables. Now it is possible to investigate the process of shape formation in the context of the relevant information formalism, as for any policy of the collective a utility function is defined. This in turn reveals the information processing that is required on average by the collective per time step to reach a desired target configuration at a certain performance level.

The task of shape formation towards a set of target configurations is episodic. That means the task ends, whenever the collective reaches a target. At the end of an episode the world state simply stays constant for all future times t, as otherwise the definition of W_t for time steps where some episodes already ended would be problematic. Strictly speaking this makes the world dynamics non-Markovian. This is circumvented, by encoding the target states into the world dynamics, that is, once the collective reaches a target state, any action leaves the world state constant. Now $P_{A|W}$ is defined to be uniform on all target states, that is $\delta(w) = 0$.

The time average distribution of the world states, denoted by the random variable W, is calculated for such a scenario as follows: The initial distribution of the world states at the beginning of an episode is given by P_{W_0}. Let T denote the uniformly distributed random variable of the time up until a time step $t_{\max}$ and L denote the binary random variable determining whether an agent is living $l = 1$ or the episode has ended $l = 0$. These depend on the actual world state random variables within the perception-action loop as illustrated in Figure 5.8. Now the marginal distribution of P_W is the limit of the future time average of the world states (for $t_{\max} \to \infty$), and the conditional distribution $P_{W|T}$ is defined as follows:

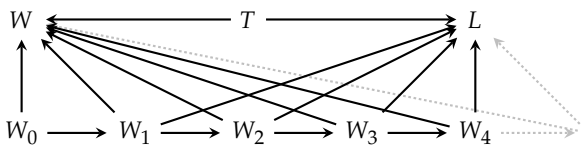


FIGURE 5.8 Illustration of the CBN of the perception-action loop of a memoryless agent with full access to the world state.

$$p(w|t) = \sum_{w_0} T^t(w, w_0)p(w_0) \tag{5.27}$$

where T is the world state transition matrix defined by $p(w'|w)$. However, I am only interested in the distribution of the world states given that the episode of an agent collective has not ended. Thus I will simply set $P_W = \lim_{t_{\max}\to\infty} P_{W|L=1}$. In practice, it is often simpler to estimate this distribution from actual simulations of episodes. However, in some cases it is desirable to determine $P_{W|L=1}$ by its analytic definition as

$$p(w|l=1) = \frac{p(l=1,w)}{p(l=1)} \tag{5.28}$$

$$= \frac{\sum_{t=0}^{t_{\max}} p(l=1,w|t)}{\sum_{t=0}^{t_{\max}} p(L=1|t)}. \tag{5.29}$$

As the world dynamics are known to ignore actions once a target was reached, the term in the sum of the denominator is simply

$$p(l=1|t) = \sum_{w} \delta(w)p(w|t). \tag{5.30}$$

Now the term in the sum in the nominator is

$$p(l=1,w|t) = p(l=1|w,t)p(w|t) = \delta(w)p(w|t). \tag{5.31}$$

And in turn it is possible to approximate P_W as the limit of $P_{W|L=1}$ for $t_{\max} \to \infty$.

In the episodic scenario, the following change of the definition of the value function is made

$$V(w) := 0 \text{ if } \delta(w) = 0 \tag{5.32}$$

i.e. the values of world states with agents in a target configuration end the recursion of the value and utility function. The value of a world state w, namely $V^\pi(w)$ is the negative of the expected number of time steps until the collective in world state w_t reaches a target configuration if the current policy is followed. This follows directly from the recursive definition of the value and utility function. Let W_0 denote the random variable of world state in the initial time step of each episode. The expected number of time steps an episode lasts is then given by

$$\bar{t}^\pi := \mathbb{E}_{W_0}[V^\pi(w_0)] \tag{5.33}$$

Thus, the total amount of information processed on average can be defined naively as

$$\mathcal{I}(p(a|w)) := I(W;A)\bar{t}^\pi. \tag{5.34}$$

The definition for agent collectives with sensors is analogously, however this definition is as remarked naive, since multiplying two averages might be different from the actual total

amount of information processed on average by an agent collective in an episode. There is a related concept called Information To-Go introduced by Tishby and Polani (2010), that directly optimizes the total information that an agent processes over the course of an episodic task, which I will only mention here as a possibly alternative to calculate the total amount of information in a information trade-off scenario.

5.7.1 *Organization in Episodic Scenarios*

If the collective follows a policy $P_{A|W}$ to form configurations specified by a target set $\tilde{X}$ it is possible that different episodes need a different number of time steps to reach one of the target configurations. Hence, it is not possible to simply calculate the multi-information of the observers, as there is in general not a time t where all possible episodes are guaranteed to have ended. In this case the organization of the collective forming a target configuration is only defined in a meaningful way in the limit of time. I will now assume that at $t = 0$ the multi-information between all location read-outs is zero, and thus the organization of the collective is as follows

$$C^{\text{org}} = \lim_{t\to\infty} I(X_t^{(1)}, ..., X_t^{(n)}). \tag{5.35}$$

The limit exists if for each state w_0 with positive probability, a target is expected to be reached in finite time, that is $\frac{\pi}{t} < \infty$, where the runtime is defined as in Eq. (5.33). This is because for any non target configuration $(x_t^{(1)}, ..., x_t^{(n)})$ and any $\varepsilon > 0$, there exists a $t > 0$ such that $p(x_t^{(1)}, ..., x_t^{(n)}) < \varepsilon$ and for all other states, namely target configurations, the dynamics of the world states are constant. That is the system if converging to a state where all world states with positive probabilities are target configurations. Hence C^{org} takes on a value that is the multi-information of an actual distribution denoted $P_{X^{(1)},...,X^{(n)}}$, i.e.

$$C^{\text{org}} = \left(\sum_{i=1}^{n} H(X^{(i)})\right) - H(X^{(1)}, ..., X^{(n)}). \tag{5.36}$$

The convergence curve of the multi-information can now provide information about speed of self-organization, as seen in Chapter 4.

5.8 SHARED CONTROL AND SENSOR COORDINATION

In Section 5.5 the following constraint was introduced for the relevant information method for agent collectives

$$p(a|w) = p(a^{(1)}|w) \cdots p(a^{(n)}|w). \tag{5.37}$$

The interpretation is, that each agent comes to its own decision concerning what action is to be taken. However a performance gain at a fixed level of information processing could be achieved by having a shared controller, that is coordination that goes beyond the knowledge of the distribution of other agents' actuators but actual correlation between

actions performed. I am not concerned with how this would be achieved and there are several possibilities: In models of biological cells, this could be achieved by a mechanism of intercellular communication or models of proto-nerves, in swarm robotics wireless communication might be possible. Here I am just interested in the theoretical limitations and possibilities that can be achieved by allowing such a shared control. Agent coordination has been considered earlier in an information-theoretic way in the context of empowerment by Capdepuy (2010).

Another part where a coordination between agents is helpful is the sensor part. This only becomes apparent if the collective does not have full world access and the exchange of sensor readings might improve the collectives performance, even in the case where each agent acts as an individual without shared control. Initial research showed, that for chains of agents, the information local sensors provide about some feature of the global configuration of the collective can be drastically improved by sharing sensor information with neighbouring agents (Harder et al., 2011).

In the following sections I will distinguish agent collectives with shared control. The policies of collectives with shared control are free from the constraint of conditional independence between the individual agent policies. In what follows, let $\bar{A} = (\bar{A}^{(1)}, ..., \bar{A}^{(n)})$ denote the joint actuator random variable of a collective with shared control and $A = (A^{(1)}, ..., A^{(n)})$ the joint actuator of a collective without shared control, as before.

5.8.1 *Intrinsic Coordination*

The amount of shared control can be measures by the intrinsic coordination of the actions, which is defined as the conditional multi-information between the agents' actuators given the world state

$$I^{\text{ic}}(\bar{A}) = I(\bar{A}^{(1)}, ..., \bar{A}^{(n)}|W). \tag{5.38}$$

It is easy to check that $I^{\text{ic}}(A)$ i.e. the intrinsic coordination of a collective without shared control, is zero. Using the relevant information mechanism it is possible to obtain two policies $\bar{\pi}$ and π for the same task, the former with shared control, the latter without. It is now possible to compare policies, that operate on the same level on information processing, that is $I(A; W) = I(\bar{A}; W)$. It is obvious that

$$\mathbb{E}[U^{\pi}(w, a)] \leq \mathbb{E}[U^{\bar{\pi}}(w, a)]. \tag{5.39}$$

If the inequality is strict the sensor information processing of the collective is more efficient with shared control. However, this does not cover all information processing per time step, as the shared control also requires an information processing of $I^{\text{ic}}(\bar{A})$ bit. Depending on the metabolic cost for this information channel, intrinsic coordination might be efficient or not, depending on the model of the

5.9 EXPERIMENTS

I will now, present some initial results I obtained in simulations with two agents, which have the task to form a bond in the center of the world. The setup consists of two agents, determined by a joint state $w = (x^{(1)}, x^{(2)}) \in \mathcal{W}$ in the state space $\mathcal{W} = \mathcal{X} \times \mathcal{X} - \Delta$ where $\mathcal{X}$ is a $w_x \times h_x$ grid-world and $\Delta = \{(x, x) | x \in \mathcal{X}\}$ the diagonal. Hence, only one agent is allowed to occupy a particular grid cell per time step. As before, the random variable representing the state of the environment is denoted by W. The goal is given by two particular adjacent cells in the centre of the grid-world and it is not relevant which agent occupies which goal cell, hence there are two goal states in the state space $\mathcal{W}$.

Each agent has five possible actions $\{N, S, W, E, H\}$, go to one of the four neighbouring cells or halt. The actions are denoted by the random variables $A^{(1)}, A^{(2)}$, and their joint action $a = (a^{(1)}, a^{(2)})$ by the random variable A as introduced in Section 5.5. The distribution of the actions only depends on the location of the two agents. In this scenario the transitions to the next step are deterministic $p(w_{t+1} | a_t, w_t) \in \{0, 1\}$ and reflect the movement of the two agent in the grid-world, blocked by the walls and blocking each other symmetrically (see Figure 5.9). The agents are blocked if they try to move to the same field or if one agent moves to a field where the other agent stays.

For every step the agents get a reward that is determined by a reward function $r(w_{t+1}, a_t, w_t)$ which depends on the current state, the action taken and the state of the world after the action was executed. A negative reward of -1 is given unless both agents occupy a goal cell in which case no reward or penalty is given.

5.9.1 *Results*

In the experiment, iterations were performed with different environment sizes (6×7,6×5,4×5, 4×3, 4×2 and $n \times 1$ with $n = 5, 6, 7, 8$). Samples were taken for different values of β ranging from 0.05 to 10.0 with steps ranging from 0.005 to 0.1, greater worlds required a larger step size due to computational limitations. Each value β leads to a policy and a state distribution, the performance of the policy can be plotted against the mutual information

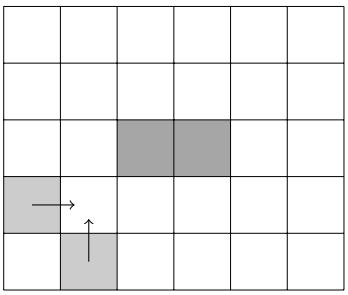


FIGURE 5.9 In this 6×5 grid-world, the two dark-grey rectangles show the goal configuration, the light-grey rectangles show a configuration where the agents block each other if they move in the directions of the arrows. This causes that the agents stay at their current position.

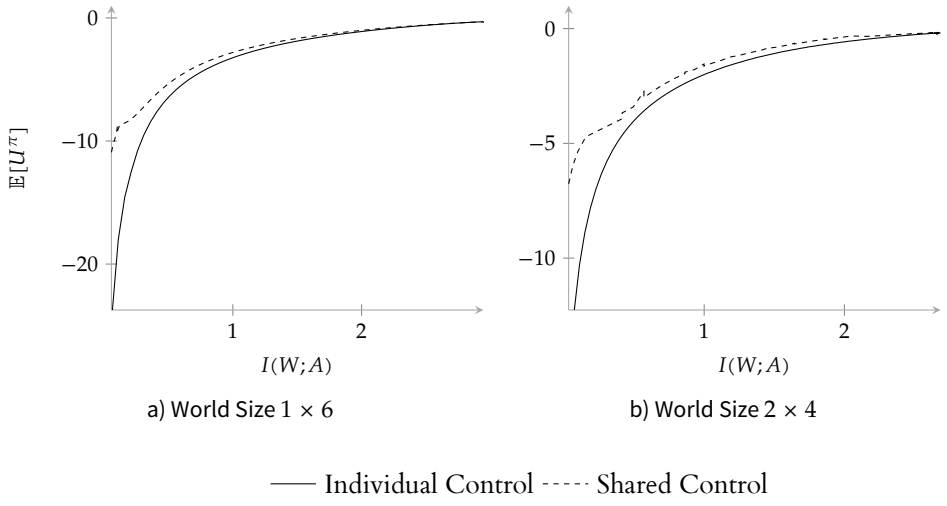


FIGURE 5.10 Performance of agents with shared controller and individual controllers with summed expectation of utility per agent and relevant information for the joint distribution of $(a^{(1)}, a^{(2)})$. Both graphs show the same features but the scales differ.

between actions and states as can be seen in Figure 5.10. At the upper limit of $\beta = 10.0$ the trade-of was already completely in favour of an optimal policy. For each sample the iteration was stopped when $\sum_s |V^\pi_{k+1}(w) - V^\pi_k(w)| < 10^{-6}$. In all runs the setup with a shared controller/policy outperforms the case where the actions are independent (see Figure 5.10). However the optimal ($\beta \to \infty$) shared controller shows almost no intrinsic coordination, that is $I^{\text{ic}}(\bar{A}^{(1)}, \bar{A}^{(2)})$ vanishes. Here the agents perform equally well with a shared controller as with independent controllers (compare Figure 5.10 and Figure 5.11). This suggests that in the optimal limit intrinsic coordination does not help to perform better. Similarly (Zahedi et al., 2009) showed that for linked robots, those performed better that had split controllers for their motors, although this was in the context of maximising predictive information.

In the suboptimal region, especially small values of β, the shared controller performs better with the same amount of relevant information. In this region the coordination behaves differently depending on the kind of controller. With independent controllers the coordination tends to zero, as less relevant information is processed (see also Figure 5.11). While this was expected due to coordination limited by relevant information, the coordination is not even close to the possible limit. The shared controller shows the opposite behaviour: the coordination increases as less relevant information is processed. This is also valid for the intrinsic coordination, which vanishes in the optimal limit.

The maximum of coordination of the shared controller depends closely on the size and geometry of the world (see Figure 5.12). The spikes in the graph are due to convergence problems for certain values of β. For larger worlds the coordination still increases for $\beta \to 0$, but by a significantly smaller amount: In a 6×7 grid world the difference between the coordination for small and large values of β is only ≈ 0.05 bit whereas in a 4×5 world

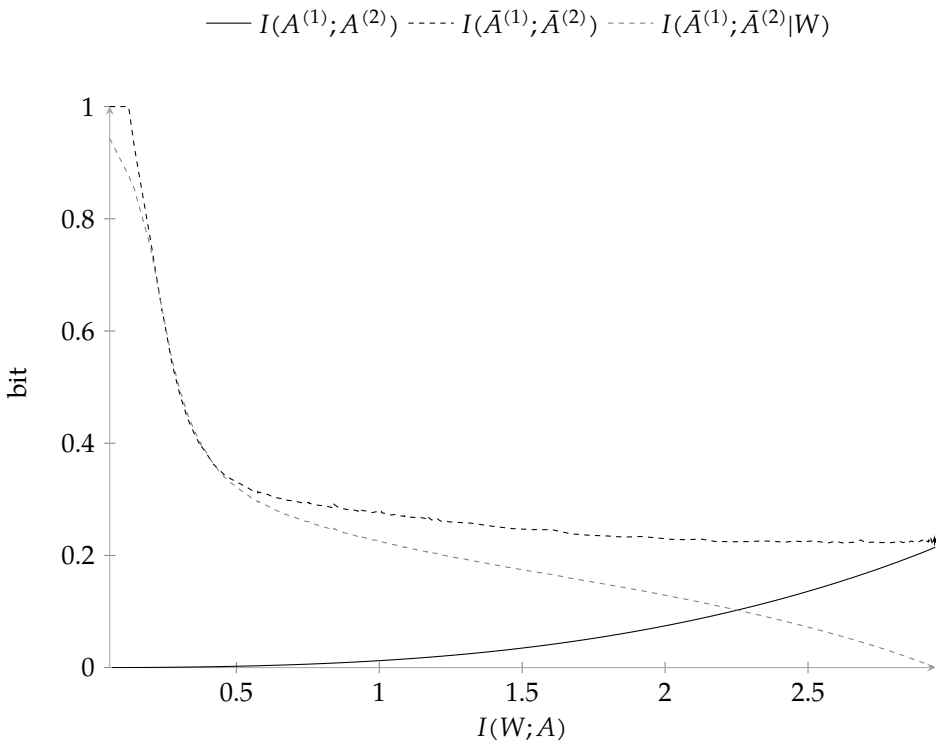


FIGURE 5.11 Coordination of agents with shared controller on a 6×1 field, comparison of intrinsic coordination for shared control $I(\bar{A}^{(1)};\bar{A}^{(2)}|W)$ with coordination for shared control $I(\bar{A}^{(1)};\bar{A}^{(2)})$ and individual control $I(A^{(1)};A^{(2)})$.

the difference is ≈ 1.54 bit. For very narrow worlds (size $n \times 1$) the coordination even reached its maximum $\max H(\bar{A}^{(1)}) = \max H(\bar{A}^{(2)}) = 1$ bit. It may seem unintuitive that this can happen while the relevant information is positive, as it means that one action fully determines the other and each of the two possible actions is chosen with probability $\frac{1}{2}$. However the coordination takes the expectation over all states: the actions can be totally synchronised, that is, $H(\bar{A}^{(1)}|\bar{A}^{(2)}) = 0$ while $H(\bar{A}^{(1)}|W)$ is not maximal. Thus the distribution of the possible two synchronous actions is not uniform, but this effect can vanish when the expectation over all states is taken, which can also be seen by that fact that the intrinsic coordination does not equal the coordination and therefore the actions cannot be independent of the states.

The distribution of the states is not uniform and W has rather low entropy as the cells that are closer to the goal are visited more often by the agents. To ensure that the observed behaviour of coordination is prevalent over the whole state space and not just appearing close to the goal the resulting policies were also analysed assuming a uniform distribution of W, which resulted only in insignificant differences.

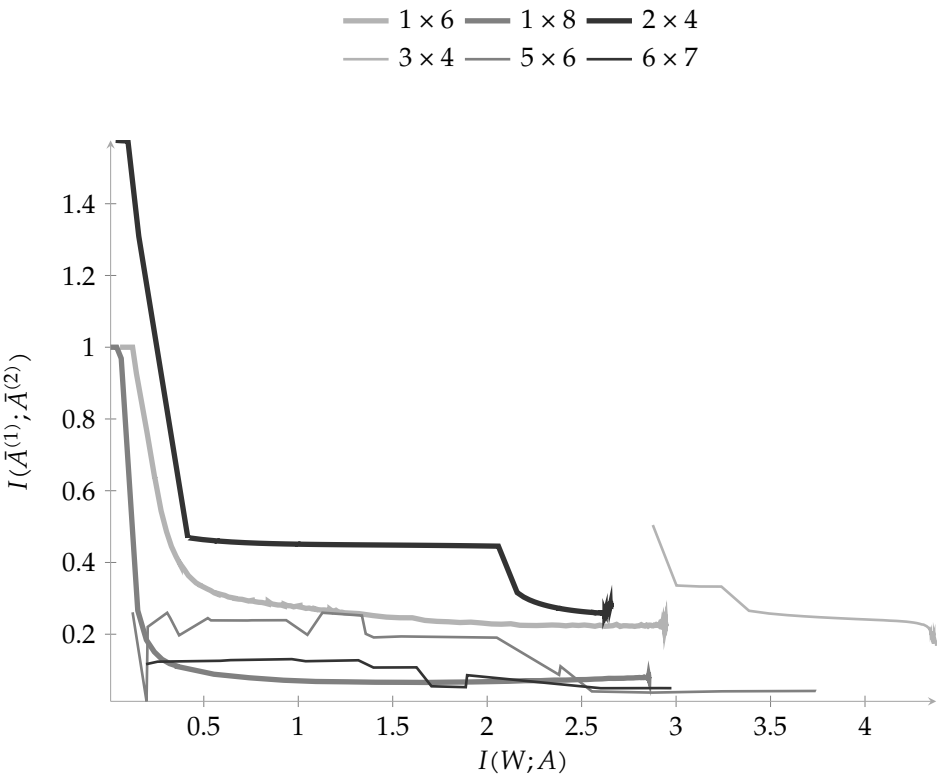


FIGURE 5.12 Coordination of agents with shared controllers in worlds of different sizes.

5.10 DISCUSSION

In this chapter I extended the concept of relevant information (Polani et al., 2006) to multi-agent systems and introduced an abstract episodic model of morphogenetic shape formation processes that can be used in conjunction with relevant information to study the information processing of collective self-organization. Moreover, I linked the information processing by individual agents to the measure of self-organization by (Polani, 2009) as introduced in Chapter 3 and introduced a formal information-theoretic definition of coordination between agents.

While the representation of embodied agents using perception-action loops was touched, this topic is still very young and there are not many results to start with, even though I propose that the transition from a simple Markov chain to a perception-action loop might be closely related to a transition from physics to biology and thus related to the origin of life.

The link between self-organization and information processing via information based control theory is very promising, but still carries at lot of assumptions with it. Possible extensions include the separation of noise and deterministic dynamics as done in (Touchette and Lloyd, 2004), which might provide a better insight into the effect of $\Delta H^{\max}_{\text{open}}$ which at

the moment includes entropy production from noise as well as entropy reduction from the dynamics of the world.

The setting in which I investigated the introduced coordination measure is a grid world with two agents and a goal to form a bond at the center of the world. As both agents have the same possible two goal states, they have to cooperate to reach the goal in an optimal way. The actions only depend on the current location of the agent (the agents are memoryless) thus the joint intent to move to the goal states is explicitly encoded in the controllers. Using an alternated fixed point iteration method I computed optimal policies for the agents under information processing constraints.

The results show that agents use intrinsic coordination to overcome limitations of their environment. This coordination is not needed in the optimal case where every agent can get all the relevant information from the environment that it needs to choose an optimal action. Though plausible, this is not entirely obvious a priori and depends on the particular task. One could think of various scenarios where the controllers are stochastic and the precise knowledge of the others agent action would lead to a better performance.

Now, large agent collectives will usually perform suboptimal policies as each agents' abilities will be limited: In real environments, the size of the agent and its supply of energy are just some limiting factors to information processing capabilities. Furthermore having many agents acting in the environment leads to spatial limitations that were here matched by the situation of narrow grid-worlds. In these cases intrinsic coordination seems to perform better than just prediction of the other agents' behaviour: The shared controller cannot be split into two independent controllers. The intrinsic coordination now also gives a measure of how strong this behaviour is.

The introduced coordination measure only gives a theoretical limit on the raw information processing capabilities of such a distributed system. In the setup above I studied the intrinsic communication is not limited and the information-theoretic limit can be reached: the two agents share a common 'brain'. But often coordination is only 'routed' through the environment: In the case of stigmergy the environment takes the role of the communication channel (Klyubin et al., 2004). Other ways of communication that have low interference with the environment like sound, hormones, neurotransmitters, morphogens, or radio signals qualify more to be modelled as intrinsic coordination, although their limited channel capacities must be considered. Examples where collectives of cells use molecular signalling, with almost no interference, to activate a certain behaviour in the whole collective (Marée and Hogeweg, 2001) could then be modelled as intrinsic coordination. With the help of interaction complexity (Ay et al., 2011) or the inference of common ancestors (Steudel and Ay, 2010) it might be possible to use the relevant information approach to devise also a network structure, maybe even hierarchies among agents that lead to a constructive approach of optimal distributed information processing in the collective.

What is still missing are implementations of the relevant information formalism for larger multi-agent systems. The currently used implementations of the algorithms for relevant information do not scale well and as it could be seen here allow only to study small systems. I hope that the advancement of information-theoretic methods for continuous random variables, including kernel based methods, will help to apply the formalism to systems with many dimensions, as in these settings continuous models are often surprisingly easier to handle than discrete models. The review of the multi-information estimators in Chapter 3 showed, that there are possibilities to employ information theory in those high dimensional settings. From a theoretical standpoint, there is not much stopping to transfer the framework presented here, to the continuous domain and during my research I started to work on the implementation of continuous relevant information algorithms with promising results obtained from a few very early tests.

REDUNDANT INFORMATION

6

» My definition of a redundancy is an air-bag in a politician's car. «

LARRY HAGMAN, Unknown

6.1 WHAT IS REDUNDANCY

Studies of synergies and redundancies have received attention in several areas including computational neuroscience (Gat and Tishby, 1999, Latham and Nirenberg, 2005, Brenner et al., 2000 and Balduzzi and Tononi, 2008), complexity sciences and genetic regulatory networks (Liang and Wang, 2008 and Margolin et al., 2006). However, there is no agreement how to best measure *redundancy* and *synergy* in an information-theoretic fashion as information is a very intricate concept. The colloquial use of the term information does not fully capture its information-theoretic meaning. This insufficiency holds also for the term redundancy and therefore the properties of redundant information can at times conflict with an intuitive, but vague feeling for what redundancy should mean. If that was not enough, it is also disputed what the formal requirements for a measure of redundant information are. I will use the term redundant information to denote information that is shared between variables (with respect to a third variable). This quantity is also sometimes denoted as *shared information* and is not to be confused with the idea of redundancy with respect to compression and entropy of an individual random variable.

My initial motivation to use a measure of redundant information and the partial information decomposition by Williams and Beer (2010) was to reach a better understanding of the morphological computation of agents in a collective. Consider a particle system as in Chapter 4, but with agents that can manipulate for example their type or act by changing the effects of interactions. The transition from one world state to the next consists of information processing performed by the agents but also by the environment. In this scenario I wanted to understand the relations between these two 'channels' of information processing. However, it became quickly apparent that the current measures are flawed according to some basic intuitions about redundancy and synergy. Hence, I set out to understand these concepts better to address these issues and will here introduce a new bivariate measure of redundant information and discuss its properties and relation to existing approaches.

Though studies on the transition of world states in the perception-action loop have not yet been giving conclusive results the measure presented here has some important applications to the study of information processing in distributed systems as for example multi-agent collectives.

6.1.1 *A Naive Approach*

Given three (finite) random variables X_1, X_2 and Z, it is possible to measure the mutual information between the joint variable (X_1, X_2) and Z denoted by $I(Z; X_1, X_2)$. The

question of redundant information is now the question of how much information that X_1 contains about Z is also contained in X_2 (about Z)? A naive answer would be that $I(Z; X_1) - I(Z; X_1|X_2)$, also called *interaction information* (Bell, 2003), measures the amount of redundant information, as it is the difference between the information about Z that is contained in X_1 and the information that is still contained in X_1 about Z if X_2 is known. However, interaction information is not sufficient to capture redundancy because it also measures synergy, whereby synergy contributes negatively to the interaction information. So, in the example of a XOR-gate $Z = X_1 \oplus X_2$ where there is no redundant information, interaction information is -1 bit. More importantly, in a situation where there is redundant and synergistic information this means that both terms interfere with each other. Therefore, the actual structure of the informational contributions is opaque to the measure of interaction information and new approaches are needed.

6.2 MEASURE CANDIDATES

As mentioned above, the term redundancy has been used in several contexts denoting different quantities. Here, I will specificly consider information about another random variable that is shared among several random variables and specificly mean the same 'piece' of information. The general setting consists of a set of finite random variables $X_{\mathbf{V}} = \{X_1, ..., X_n\}$, the index set $\mathbf{V} = \{1, ..., n\}$ and a finite random variable Z with values from $\mathcal{X}_1 \times ... \times \mathcal{X}_n$ and $\mathcal{Z}$ respectively. The mutual information between Z and X_V is denoted as follows:

$$I(Z; X_{\mathbf{V}}) := I(Z; X_1, ..., X_n). \tag{6.1}$$

In the bivariate setting this contracts simply to $I(Z; X_1, X_2)$. The main question in this chapter is how to quantify redundancy between the variables X_i with respect to Z and in turn get a deeper insight into the structure of mutual information. For this, I will be using a framework introduced by Williams and Beer (2010). Here, an introduction to existing approaches to redundant information will be given including a comparison with the newly constructed measure.

6.2.1 *Minimal Information*

The first candidate measure for redundant information I consider is called *minimal information*. In the following it is denoted by $I_{\min}$ (Williams and Beer, 2010). Following Williams and Beer (2010), the construction of the measure starts by considering the (non-negative) *specific information* (DeWeese and Meister, 1999), which is the increase in likelihood (or reduction in surprise) of the outcome of a specific event $z \in \mathcal{Z}$ and $x_{\mathbf{A}}$ with respect to an index set $\mathbf{A} \subseteq \mathbf{V}$ and is defined by

$$I_{\text{sp}}(Z = z; \mathbf{A}) := \sum_{x_{\mathbf{A}}} p(x_{\mathbf{A}}|z) \left[\log \frac{1}{p(z)} - \log \frac{1}{p(z|x_{\mathbf{A}})} \right] \tag{6.2}$$

$$= D_{\text{KL}} \left(P_{X_{\mathbf{A}}|Z} \middle\| P_{X_{\mathbf{A}}} \right) \tag{6.3}$$

where the equality results from applying Bayes' rule. This definition is in turn used to define the *minimal information* that a set of (joint) random variables contains about the outcome (Williams and Beer, 2010) as

$$I_{\min}(Z; \mathbf{A}_1, ..., \mathbf{A}_k) := \sum_z p(z) \min_i I_{\text{sp}}(Z = z; \mathbf{A}_i). \tag{6.4}$$

Minimal information is obviously non-negative and, in fact, positive if all variables X_{A_i} with respect to the index sets A_i contain some information about a specific outcome (for outcomes having positive probabilities).

The notation deviates from many information theoretic measures as the parameters of $I_{\min}$ are index sets and not random variables. Nonetheless, I adopt this notation from Williams and Beer (2010) for multivariate measures of redundant information to make comparisons between the measures easier to comprehend. For the bivariate case the notation changes slightly and I will use the random variables directly instead of the index set notation, so instead of writing $I_{\min}(Z; A_1, A_2)$, where A_1 and A_2 are index sets of some collection of random variables, I will simply write $I_{\min}(Z; X_1, X_2)$.

6.2.2 *Comparing Apples & Oranges*

This measure contradicts a basic intuition about redundancy. Consider the case of two binary input variables X_1, X_2 (i.e. $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$) that are independent, uniformly distributed and where $Z = (X_1, X_2)$ is an unaltered copy of both variables, i.e. the underlying distribution of Z is the joint distribution of X_1 and X_2. Now it is to be expected that there is be no redundancy between X_1 and X_2 with regard to Z because X_1 and X_2 are independent, so the information contained about Z in X_1 and X_2 respectively is clearly not the same. However, an easy calculation leads to $I_{\min}(Z; X_1, X_2) = 1$ bit.

Minimal information information is observed because for each outcome of X_1 or X_2 a reduction of entropy regarding an outcome z is observed (i.e. the specific information between X_1 and z as well as X_2 and z is positive). This ignores that even though X_1 and X_2 give the same amount of information about an outcome z, they tell something different about the change of the distribution P_Z. In this particular example X_1 gives information about the first component of Z while, X_2 gives information about the second component of Z. This example is used to demonstrate the effect with full impact, though measuring a larger minimal information than what is considered to be redundant can also occur in more practical situations. Whenever there is a process that has independent sub-components over time and these components contain some information about their future states, the measure $I_{\min}$ will report this information as redundancy between the components.

More precisely the *a posteriori* distributions of Z, $P_{Z|X_1}$ and $P_{Z|X_2}$, when either X_1 or X_2 have been observed, give a different *kind* of information (have different content) even though they give the same *amount* of information. The core idea to resolve this issue therefore is to separate the contributions of X_1 and X_2 by adopting a geometric view in the space of probability distributions over Z.

6.2.3 *Axiomatic Approach*

Before looking into other proposed measures, I will describe an axiomatic approach that will be considered for all measures introduced from here on. Williams (2011) states three axioms any redundancy measure $I_\cap(Z; A_1, ..., A_k)$ has to fulfil:

Weak Symmetry (S_0) $I_\cap$ is symmetric with respect to permutations of the A_i's.

Self-Redundancy (I) $I_\cap(Z; A) = I(Z; X_A)$.

Monotonicity (M) $I_\cap(Z; A_1, ..., A_{k-1}, A_k) \leq I_\cap(Z; A_1, ..., A_{k-1})$ with equality if $A_{k-1} \subseteq A_k$.

These axioms follow the intuition that redundancy with regard to a variable is symmetric with respect to permutations of the input variables and similar to how entropy can be viewed as self-information, i.e. $H(X) = I(X; X)$, mutual-information can be viewed as self-redundancy $I(Z; X_A) = I_\cap(Z; A)$. The last axiom is also intuitive, considering that redundancy denotes information about Z that is contained in every variable X_{A_i}, each additional variable is a further constraint, so that the redundancy can only be reduced. The only exception is where the additional variable is a joint variable of an already used variable and any arbitrary other random variable, in this case the redundancy stays constant.

From these axioms follows the non-negativity of the redundancy measure, as well as that it is bounded from above by the mutual information between Z and each source. To prove this, note that A_i are subsets of V that could be empty, and for consistency $I_\cap(Z; \emptyset) = 0$ by definition. It is easy to check that all three axioms are fulfilled by the measure $I_{\min}$ (Williams, 2011).

6.2.3.1 *Identity Axiom*

To address the shortcoming of the minimal information which was identified above, I propose to add an additional axiom to the axioms introduced above and call it the *identity property*, as it states how redundancy should behave with respect to a joint random variable of identical copies of the two source variables. It requires that for any redundancy measure $I_\cap$ the following axiom holds:

Identity (Id_2) $I_\cap\left((X_{A_1}, X_{A_2}); A_1, A_2\right) = I(X_{A_1}; X_{A_2})$

The idea behind this additional axiom is, that if the (bivariate) mechanism that is considered is just copying the input, the redundancy must be exactly the mutual information between the variables. Given a multivariate redundancy measure the monotonicity automatically states that the multivariate redundancy is then bounded from above by the minimum of pairwise mutual information terms. Later on I will discuss this property in more detail, as there is also a point of view, that does neither agree with the identity axiom nor with the redundancy calculated via $I_{\min}$.

6.2.4 *Synergistic Mutual Information*

An approach coming from cryptography is to use *intrinsic conditional mutual information* (ICMI) to measure the unique contributions to mutual information (Maurer and Wolf, 1999). These unique contributions could in theory be used to define a redundancy measure using the structure of mutual information as it will be introduced in Section 6.4.1. However, as noted by Bertschinger et al. (2012), the ICMI does not obey the consistency of the partial information decomposition which will be introduced later in Section 6.4 and it is fair to say that this measure does not add up for this use case. Therefore, I will not go into the details of its construction here.

However, there is a measure of synergistic information by Griffith and Koch (2012), denoted *synergistic mutual information*, whose construction is inspired by the construction of ICMI. Again, using the structure of mutual information it is possible to deduce a measure of redundant information from this.

The measure first defines *union information*, denoted $I_\cup$, as the amount of information in the individual X_{A_i} about Z, but without counting the same 'piece' of information twice. This is achieved as follows:

$$I_\cup(Z; X_{A_1}, \ldots, X_{A_k}) := \min_{\substack{p(\tilde{z}|z) \\ (X_{A_1},\ldots,X_{A_k}) \to Z \to \overline{Z} \\ \forall i: I(X_i;\overline{Z}) = I(X_i;Z)}} I(\overline{Z}; X_{A_1}, \ldots, X_{A_k}), \tag{6.5}$$

where $\overline{Z}$ is a truncated version of Z of same cardinality, similar to the bottleneck variable in an information bottleneck (Tishby et al., 1999). The minimization constraint denotes that $(X_{A_1}, \ldots, X_{A_k}) \to Z \to \overline{Z}$ forms a Markov chain, where the information between each individual variable and Z is preserved in $\overline{Z}$. The synergistic information is now defined as

$$S(Z; X_{A_1}, \ldots, X_{A_k}) := I(Z; X_{A_1}, \ldots, X_{A_k}) - I_\cup(Z; X_{A_1}, \ldots, X_{A_k}). \tag{6.6}$$

Again, it is possible to use this quantity to define a corresponding redundancy measure. However, as I will discuss Section 6.5.2, this measure does not allow a consistent decomposition.

6.2.5 *Shared Information*

A recent article by Bertschinger et al. (2012) extends the discussion about redundancy, or as it is called in the article, *shared information*. In the article the authors introduce three new axioms or properties that they argue are necessary for a measure of redundant information. In comparison to the axioms mentioned above, the left side variable Z plays a role in these properties:

Strong Symmetry (S_1) $I_\cap$ is symmetric with respect to permutations of Z and the A_i's.

Left Monotonicity (LM) $I_\cap(Z; A_1, ..., A_{k-1}, A_k) \leq I_\cap(Z, Z'; A_1, ..., A_{k-1}, A_k)$.

Left Chain Rule (LC) $I_\cap(Z, Z'; A_1, ..., A_{k-1}, A_k) = I_\cap(Z; A_1, ..., A_{k-1}, A_k) + I_\cap(Z'; A_1, ..., A_{k-1}, A_k|Z)$.

They go on and show that there are certain subsets of these axioms that cannot be fulfilled by any measure, which obviously is an important result and already gives an additional insight into the structure of mutual information. The article continues by defining two measures of shared information, but both having quite drastic shortcomings. I postpone a comparison of all measures to Section 6.5 where my newly proposed measure and also, very importantly, the decompositional structure of mutual information will have been introduced.

6.3 CONSTRUCTION OF A NEW MEASURE

To define a new (bivariate) redundancy measure I will take a geometric view on informational quantities. Information geometry is a powerful tool-set to investigate information-theoretic question in the context of Riemannian manifolds (Amari, 2001 and Amari and Nagaoka, 2007). Geometric arguments and algorithms have profound application to information theory and statistics (Csiszar and Shields, 2004), and have been successfully employed to construct information-theoretic multivariate interaction measures (Kahle et al., 2009). Information geometry deals with statistical manifolds of probability distributions equipped with the Fisher metric (Amari and Nagaoka, 2007). The Kullback-Leibler divergence is now a divergence function on the statistical manifold and thus certain helpful properties and theorems, such as the Pythagorean Theorem, can be used. Here, I will introduce concepts of information geometry only as needed, because most arguments can be done on an ad-hoc basis.

6.3.1 *Preliminaries*

The redundancy measure that is constructed in the subsequent sections is based on the notion of *projected information* which I will introduce shortly. In what follows, let $\Delta(Z)$

denote the space of all probability distributions over Z. An information projection is now defined as the minimization of the Kullback-Leibler divergence between a probability distribution in $P \in \Delta(Z)$ and a subset $B \subset \Delta(Z)$:

$$\pi_B(P) := \operatorname*{arg\,min}_{R \in B} D_{\mathrm{KL}}(P \| R). \tag{6.7}$$

The Kullback-Leibler divergence is not symmetric, therefore it is possible to define a dual projection $\pi_B{}^*(P)$ where the parameters of $D_{\mathrm{KL}}(\cdot \| \cdot)$ are reversed (in (Csiszár and Matus, 2003), $\pi_B(P)$ is called reverse information projection and $\pi_B{}^*(P)$ information projection). Here I will exclusively use the projection $\pi_B(P)$.

For $B \subseteq \Delta(Z)$, let

$$C_{\mathrm{cl}}(B) = \{\lambda P + (1-\lambda)Q | P, Q \in B, \lambda \in [0,1]\} \tag{6.8}$$

denote the convex closure of B in $\Delta(Z)$. As $\Delta(Z)$ is convex it follows that $C_{\mathrm{cl}}(B) \subseteq \Delta(Z)$. Observing an event x_1 in X_1 or x_2 in X_2 leads to a distribution over Z, $P_{Z|x_1} \in \Delta(Z)$ and $P_{Z|x_2} \in \Delta(Z)$ respectively. Let

$$\langle X_1 \rangle_Z := \{P_{Z|x_1} : x_1 \in \mathcal{X}_1\} \tag{6.9}$$

denote the set of all conditional distributions of Z for the different events of X_1. Because the marginal distributions over Z are a convex combination of the conditional distributions, namely

$$p(z) = \sum_{x_1} p(z|x_1)p(x_1), \tag{6.10}$$

the space of distributions over X_1, i.e. $\Delta(X_1)$, is embedded in $\Delta(Z)$ in the following way

$$C_{\mathrm{cl}}(\langle X_1 \rangle_Z) = C_{\mathrm{cl}}\left(\{P_{Z|x_1} : x_1 \in \mathcal{X}_1\}\right) \tag{6.11}$$

and thus $C_{\mathrm{cl}}(\langle X_1 \rangle_Z) \subseteq \Delta(Z)$. Assuming that the mechanism $P_{Z|x_1}$ is known for all x_1, the convex closure of $\langle X_1 \rangle_Z$ in $\Delta(Z)$ now contains all marginals P_Z that could be the actual marginal of Z if the underlying distribution of X_1 is not known. Conversely, for each $P_Z \in C_{\mathrm{cl}}(\langle X_1 \rangle_Z)$ there is a way to represent P_Z as a convex combination of the distributions $P_{Z|x_1}$ (because $C_{\mathrm{cl}}(\langle X_1 \rangle_Z)$ is a convex closure of a finite set of points), the coefficients of the convex combination are then the probabilities $p(x_1)$.

For example, the problem of finding the maximal channel capacity between two random variables X_1 and Z, with a given input distrubtion on X_1 and Z as output, can now be rephrased as finding the point P_Z in the convex closure $C_{\mathrm{cl}}(\langle X_1 \rangle_Z)$ that maximizes its Kullback-Leibler divergence from all extremal points $P_{Z|x_1}$ of the convex closure (weighted by the respective probabilities $p(x_1)$), i.e. that maximizes

$$I(X_1; Z) = \sum_{x_1} p(x_1) D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_Z\right). \tag{6.12}$$

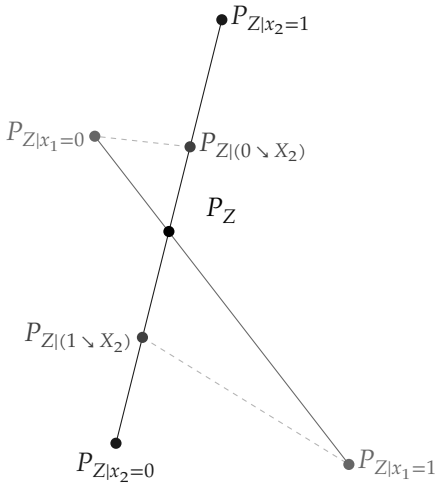


FIGURE 6.1 Illustration of the construction of projective information for binary input variables. Points represent the distributions in the space of distributions over the variable Z. The lines connecting points denote the subspace of conditional distributions depending on the distribution of X_1 and X_2 respectively.

6.3.2 *Projective Information*

Information projections can now project the conditionals of one variable onto the convex closure of the other. I will denote this projection by

$$P_{Z|(x_1 \searrow X_2)} := \pi_{C_{\mathrm{cl}}(\langle X_2 \rangle_Z)}(P_{Z|x_1}). \tag{6.13}$$

The projection is not guaranteed to be unique (for uniqueness, the set onto which is projected would need to be log-convex and not convex (Csiszár and Matus, 2003)), however this does not matter for my purposes as can be seen in the next lemma.

Now, the *projected information* of X_1 onto X_2 with respect to Z is defined as

$$I_Z^\pi (X_1 \searrow X_2) := \sum_{z,x_1} p(z, x_1) \log \frac{p_{Z|(x_1 \searrow X_2)}(z)}{p(z)}. \tag{6.14}$$

The rationale behind this construction is that the projected information quantifies the amount of information that two variables share with each other, here X_1 and Z, that can be expressed in terms of the information X_2 shared with Z (projecting onto X_2). This is illustrated for binary input variables Figure 6.1.

Lemma 6.1. *Projected information $I_Z^\pi (X_1 \searrow X_2)$ is well-defined, finite and non-negative.*

Proof. First, note that projected information can be written as the difference of two Kullback-Leibler divergences

$$\begin{aligned} I_Z^\pi (X_1 \searrow X_2) = \sum_{x_1} p(x_1) \Big[& D_{\mathrm{KL}} \left(P_{Z|x_1} \middle\| P_Z \right) \\ & - D_{\mathrm{KL}} \left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)} \right) \Big]. \end{aligned} \tag{6.15}$$

Therefore, if the projection is not unique, projected information only takes the KL-divergence into account which is the same for all possible solutions of the minimization problem in (6.7). Now $D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right) \leq D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_Z\right)$ for all $x_1 \in \mathcal{X}_1$ because $P_Z \in C_{\mathrm{cl}}(\langle X_2 \rangle_Z)$ and the definition of $P_{Z|(x_1 \searrow X_2)}$ as the distance minimizing distribution to $P_{Z|x_1}$ in $C_{\mathrm{cl}}(\langle X_2 \rangle_Z)$. Hence $I_Z^\pi\left(X_1 \searrow X_2\right) \geq 0$. Furthermore

$$I(X_1; Z) = \sum_{x_1} p(x_1) D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_Z\right) < \infty. \tag{6.16}$$

6.3.3 *Definition of Bivariate Redundancy*

The (bivariate) redundancy measure is now simply defined as the minimum of both projected information terms

$$I_{\mathrm{red}}(Z; X_1, X_2) := \min\{I_Z^\pi\left(X_1 \searrow X_2\right), I_Z^\pi\left(X_2 \searrow X_1\right)\}. \tag{6.17}$$

At this point it is possible to take the minimum over both values because the values are already corrected for the change of the distributions in different directions by projecting the conditionals. This definition is different to the approach taken by Williams and Beer (2010), where the minimization does not consider that events in different source variables may change the distribution of the outcome in different directions in the geometrical space of distributions. Self-redundancy is now explicitly defined as

$$I_{\mathrm{red}}(Z; X_1) := I_{\mathrm{red}}(Z; X_1, X_1) = I_Z^\pi\left(X_1 \searrow X_1\right) \tag{6.18}$$

to fulfil the redundancy axioms.

6.3.4 *The Proposed Measure is a Bivariate Redundancy Measure*

To show that this is actually a redundancy measure, it needs to be shown that it fulfils the four axioms (weak-symmetry, self-redundancy, monotonicity and identity). Weak-symmetry is obviously fulfilled, self-redundancy is also very quick to prove:

$$\begin{aligned} I_{\mathrm{red}}(Z; X_1) &= I_Z^\pi\left(X_1 \searrow X_1\right) && (6.19)\\ &= \sum_{z, x_1} p(z, x_1) \log \frac{p_{Z|(x_1 \searrow X_1)}(z)}{p(z)} && (6.20)\\ &= \sum_{z, x_1} p(z, x_1) \log \frac{p(z|x_1)}{p(z)} && (6.21)\\ &= I(Z; X_1). && (6.22) \end{aligned}$$

The inequality part of the monotonicity axiom is directly given by the following proposition.

Proposition 6.1. $I_{\mathrm{red}}(Z; X_1, X_2) \leq I(Z; X_1)$

Proof. Using the expression of projected information as a difference of Kullback-Leibler divergences leads to

$$
\begin{aligned}
I_{\mathrm{red}}(Z; X_1, X_2) &\leq I_Z^{\pi}\left(X_1 \searrow X_2\right) && (6.23)\\
&= \sum_{x_1} p(x_1) \Big[D_{\mathrm{KL}}\left(P_{Z|X_1} \middle\| P_Z\right) \\
&\quad - D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right) \Big] && (6.24)\\
&= I(Z; X_1) \\
&\quad - \sum_{x_1} p(x_1) D_{\mathrm{KL}}\left(P_{Z|X_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right). && (6.25)
\end{aligned}
$$

Hence it follows that $I_{\mathrm{red}}(Z; X_1, X_2) \leq I(Z; X_1)$ as the KL-divergence is non-negative (Cover and Thomas, 2006). □

The following is needed to show equality holds if $X_2 = (X_1, X_3)$, where X_3 is an arbitrary finite random variable.

Lemma 6.2. *For all $x_1 \in \mathcal{X}_1$ and random variables X_2 and X_3,*

$$
\sum p(z|x_1) \left(\log p_{Z|(x_1 \searrow (X_2, X_3))}(z) - \log p_{Z|(x_1 \searrow X_2)}(z) \right) \geq 0. \tag{6.26}
$$

Proof. Let $x_1 \in \mathcal{X}_1$, as $C_{\mathrm{cl}}(\langle X_2 \rangle_Z) \subseteq C_{\mathrm{cl}}(\langle (X_2, X_3) \rangle_Z)$ (note that $p(z|x_2) = \sum_{x_3} p(x_3|x_2) p(z|x_2, x_3)$) it follows due to the definition of the projection that

$$
\begin{aligned}
&\sum p(z|x_1) \log \frac{p(z|x_1)}{p_{Z|(x_1 \searrow (X_2, X_3))}(z)} \leq \sum p(z|x_1) \log \frac{p(z|x_1)}{p_{Z|(x_1 \searrow X_2)}(z)} && (6.27)\\
\iff &\sum p(z|x_1) \log p_{Z|(x_1 \searrow (X_2, X_3))}(z) \geq \sum p(z|x_1) \log p_{Z|(x_1 \searrow X_2)}(z) && (6.28)
\end{aligned}
$$

□

Lemma 6.3. *For all $(x_2, x_3) \in \mathcal{X}_2 \times \mathcal{X}_3$*

$$
\sum p(z|x_2, x_3) \left(\log p_{Z|((x_2, x_3) \searrow X_1)}(z) - \log p_{Z|(x_2 \searrow X_1)}(z) \right) \geq 0. \tag{6.29}
$$

Proof. Here $R = P_{Z|((x_2, x_3) \searrow X_1)}$ is minimizing $D_{\mathrm{KL}}\left(P_{Z|x_2, x_3} \middle\| R\right)$ by definition, therefore

$$
\begin{aligned}
&\sum p(z|x_2, x_3) \log \frac{p(z|x_2, x_3)}{p_{Z|((x_2, x_3) \searrow X_1)}(z)} \leq \sum p(z|x_2, x_3) \log \frac{p(z|x_2, x_3)}{p_{Z|(x_2 \searrow X_1)}(z)} \\
\iff &\sum p(z|x_2, x_3) \log p_{Z|((x_2, x_3) \searrow X_1)}(z) \geq \sum p(z|x_2, x_3) \log p_{Z|(x_2 \searrow X_1)}(z)
\end{aligned}
$$

□

Proposition 6.2. $I_{\mathrm{red}}(Z; X_1, X_2) \leq I_{\mathrm{red}}(Z; X_1, (X_2, X_3))$

Proof. From Lemma 6.2 it follows directly that $I_Z^\pi(X_1 \searrow X_2) \leq I_Z^\pi(X_1 \searrow (X_2, X_3))$, furthermore from Lemma 6.3, $I_Z^\pi(X_2 \searrow X_1) \leq I_Z^\pi((X_2, X_3) \searrow X_1)$ respectively. Hence the conclusion $I_{\text{red}}(Z; X_1, X_2) \leq I_{\text{red}}(Z; X_1, (X_2, X_3))$. □

Proposition 6.1 states that $I_{\text{red}}(Z; X_1, X_2) \leq I(Z; X_1)$ and thus for $X_2 = (X_1, X_3)$ also $I_{\text{red}}(Z; X_1, (X_1, X_3)) \leq I(Z; X_1)$, the proposition above now also proves that the inequality in the other direction also holds

$$\begin{aligned} I(Z; X_1) &= I_{\text{red}}(Z; X_1) && (6.30)\\ &= I_{\text{red}}(Z; X_1, X_1) && (6.31)\\ &\leq I_{\text{red}}(Z; X_1, (X_1, X_3)). && (6.32) \end{aligned}$$

Hence, the equality case of the monotonicity **(M)** holds.

Now it is only left to show that the measure also fulfils the new identity property **($\mathbf{Id_2}$)**, namely

$$I_{\text{red}}((X_1, X_2); X_1, X_2) = I(X_1; X_2). \tag{6.33}$$

To prove the identity property the following technical lemma is needed

Lemma 6.4. *If $Z = (X_1, X_2)$ and (x_1', x_2') denotes an event of Z then $p_{Z|(x_2' \searrow X_1)}(x_1', x_2') = p_{Z|(x_1' \searrow X_2)}(x_1', x_2') = p(x_1'|x_2')p(x_2'|x_1')$.*

Proof. Let $R \in C_{\text{cl}}(\langle X_1 \rangle_Z)$, it is of the form

$$r(x_1', x_2') = \sum_x \alpha_x p(x_1', x_2'|x_1) = \alpha_{x_1'} p(x_2'|x_1'), \tag{6.34}$$

where $\alpha_x \geq 0$ and $\sum \alpha_x = 1$. This means that any distribution of X_1 is embedded in (X_1, X_2) by scaling with the conditional distribution $P_{X_2|X_1}$, which is nothing else than basic probability calculus with $\alpha_{x_1'} = p(x_1')$ representing the embedded distribution. Now let

$$\begin{aligned} L_{x_2}(\alpha_{x_1'}) &:= D_{\text{KL}}\left(P_{Z|x_2} \middle\| R\right) && (6.35)\\ &= \sum_{x_1', x_2'} p(x_1', x_2'|x_2) \log \frac{p(x_1', x_2'|x_2)}{\alpha_{x_1'} p(x_2'|x_1')} && (6.36)\\ &= \sum_{x_1'} p(x_1'|x_2) \log \frac{p(x_1'|x_2)}{\alpha_{x_1'} p(x_2|x_1')}. && (6.37) \end{aligned}$$

In Eq. (6.36) the definition of the KL-divergence is used and r is replaced with its form from Eq. (6.34). The next step takes into account that $p(x_1', x_2'|x_2) = 0$ if $x_2' \neq x_2$, and thus it is possible to replace x_2' by x_2 throughout the term. A simple, but tedious, and therefore here omitted calculation shows now that the point at $\alpha_{x_1'} = p(x_1'|x_2)$ fulfils the optimality conditions of Karush-Kuhn-Tucker (KKT) (Kuhn and Tucker, 1951) for the

minimization of $L_{x_2}(\alpha_{x_1'})$ with the simplex constraint of $\alpha_{x_1'}$. The KL-divergence is convex in the second parameter (Cover and Thomas, 2006) and thus it follows from the KKT conditions (Boyd and Vandenberghe, 2004) that $\alpha_{x_1'} = p(x_1'|x_2)$ is a global solution for the constrained minimization of $L(\alpha_{x_1'})$ and in turn, the constrained minimization of the KL-divergence $D_{\mathrm{KL}}(p(Z|x_2) \,\|\, r)$ gives $r(x_1', x_2') = p(x_1'|x_2)p(x_2'|x_1')$.

Now for the proof of the lemma, the projected distribution $P_{X_1', X_2'|(x_2' \searrow X_1)}$, which is evaluated at (x_1', x_2') is not using an arbitrary $P_{Z|x_2}$ as it was done so far in this proof, but specifically $P_{Z|x_2'}$. Therefore it is possible to set $x_2 = x_2'$, so that $r(x_1', x_2') = p(x_1'|x_2')p(x_2'|x_1')$ and it follows that $p_{Z|(x_2' \searrow X_1)}(x_1', x_2') = p(x_1'|x_2')p(x_2'|x_1')$. The converse $p_{Z|(x_1' \searrow X_2)}(x_1', x_2') = p(x_1'|x_2')p(x_2'|x_1')$ is shown analogously. □

Hence the proof of Lemma 6.4 concludes with the following proposition:

Proposition 6.3. $I^{\pi}_{X_1,X_2}(X_1 \searrow X_2) = I^{\pi}_{X_1,X_2}(X_2 \searrow X_1) = I(X_1; X_2)$

Proof. Without loss of generality,

$$I^{\pi}_{X_1,X_2}(X_1 \searrow X_2) \tag{6.38}$$

$$= \sum_{x_1', x_2', x_1} p(x_1', x_2', x_1) \log \frac{p_{X_1,X_2|(x \searrow X_2)}(x_1', x_2')}{p(x_1', x_2')} \tag{6.39}$$

$$= H(X_1, X_2) + \sum_{x_1', x_2'} p(x_1', x_2') \log p_{X_1,X_2|(x_1' \searrow X_2)}(x_1', x_2') \tag{6.40}$$

$$= H(X_1, X_2) + \sum_{x_1, x_2} p(x_1, x_2) \log[p(x_1|x_2)p(x_2|x_1)] \tag{6.41}$$

$$= H(X_1, X_2) - H(X_1|X_2) - H(X_2|X_1) \tag{6.42}$$

$$= I(X_1; X_2). \tag{6.43}$$

□

Thus I_{red} is a good candidate for measuring redundancy (in terms of bivariate redundancy with respect to some target variable).

6.4 PARTIAL INFORMATION DECOMPOSITION

So far, I only considered redundant information, especially in its bivariate version. However, there is the broader framework of *partial information decomposition* introduced by Williams and Beer (2010) within which redundancy plays a key role. The motivation for the partial information decomposition of multivariate mutual information (in the form of $I(Z; X_1, ..., X_n)$) is a better insight into its structure and the informational contributions of the input variables X_1,...,X_n to the outcome of Z. This also includes a measure for a complementary concept of redundancy called *synergy*, which denotes information that is only available if the outcome of several input variables is known. One of the important

property of the decomposition given a compliant redundancy measure is the non-negativity of all its terms. I will now introduce the partial information (PI) decomposition for multivariate mutual information (Williams and Beer, 2010) in further detail and, after a short excursion into multivariate terrain, show that, for the bivariate case, the proposed measure of redundant information also leads to a positive decomposition.

6.4.1 *The Structure of Mutual Information*

Let $I_\cap$ denote an arbitary redundancy measure fullfilling the axioms from Section 6.2.3. This redundancy measure $I_\cap$ is is now used to construct *partial information atoms* (PI-atoms) which measure the contributions of sets of random variables to a multivariate mutual information term. For simplicity I present the bivariate case first: From the axioms it is clear, that the redundancy $I_\cap(Z; X_1, X_2)$ is less than (or equal to) the mutual information $I(Z; X_1, X_2)$. Or put simply, the redundant information about Z in each variable is less than (or equal) the overall information X_1 and X_2 contain about Z. The redundant information is also less than the information each individual variable contains about Z, i.e. $I_\cap(Z; X_1, X_2) \leq I(Z; X)$ and $I_\cap(Z; X_1, X_2) \leq I(Z; X_2)$. Using these quantities it is possible to capture three further quantities: The *unique information* of X_1 (and X_2 respectively) which denotes the information about Z that is exclusively contained in X_1, and not shared redundantly with X_2. Finally, there is *synergistic information* which denotes all information that X_1 and X_2 contain about Z, but which is neither available individually in X_1 nor X_2. The classical example for this case is a XOR-gate, where knowing the state of one of the inputs does not give any more certainty about the state of the output, whereas knowing both inputs determines the output state without uncertainty. The decomposition of the mutual information $I(Z; X_1, X_2)$ into redundant, unique and synergistic information is illustrated in Figure 6.2

Given more random variables, this concept can be extended to decompose mutual information terms of the form

$$I(Z; X_1, ..., X_n), \tag{6.44}$$

and it is possible to ask what the redundant and synergistic contributions between any subsets of source random variables X_1,...,X_n are. As in Section 6.2.1 the set of all source random variables is denoted by $X_{\mathbf{V}} = \{X_1, ..., X_n\}$ where $\mathbf{V}$ is the index set. The redundancy measure $I_\cap$ can now be used to quantify the redundant information between several random variables. Specificly the redundancy between joint variables of index subsets of $\mathbf{V}$ is considered. For example if $\mathbf{V} = \{1, 2, 3, 4, 5\}$, $\mathbf{A} = \{1, 2, 3\}$, $\mathbf{B} = \{1, 4\}$ and $\mathbf{C} = \{5\}$ then the term

$$I_\cap(Z; \mathbf{A}, \mathbf{B}, \mathbf{C}) \tag{6.45}$$

denotes the redundancy with respect to Z between the three random variables (X_1, X_2, X_3), (X_1, X_4) and (X_5). To shorten notation I will follow the notation of Williams and Beer (2010) and drop the commata

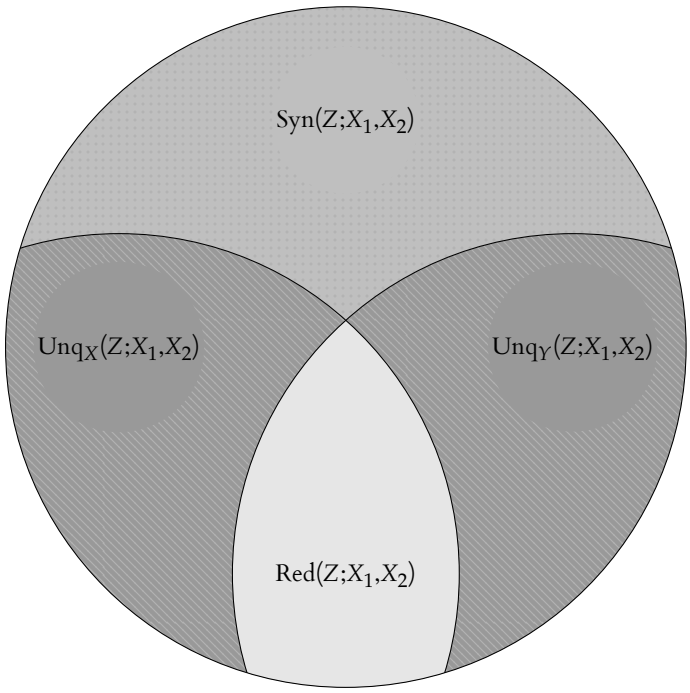


FIGURE 6.2 Illustration of the Partial Information Decomposition into redundant, unique and synergistic terms.

$$I_\cap(Z;\{123\}\{14\}\{5\}) \tag{6.46}$$

to denote $I_\cap(Z;\{1,2,3\},\{1,4\},\{5\})$ and $I_\cap(Z;\mathbf{A},\mathbf{B},\mathbf{C})$ respectively.

6.4.1.1 *Anti-Chains and The Redundancy Lattice*

Now, the idea is to look at all possible redundancies between sets of source variables. Here the monotonicity axiom (M) comes in handy, as it constrains the possibilities that need to be considered to all sets of sets of source random variables where no two sets of variables share a sub-/superset relation. For example, by the monotonicity axiom, the term $I_\cap(Z;\{123\}\{12\}\{4\})$ is equal to $I_\cap(Z;\{12\}\{4\})$ because $\{12\}$ is a subset of $\{123\}$. Formally, all these sets are determined as follows

$$\mathcal{A}(\mathbf{V}) = \{\alpha \in \mathcal{P}_1(\mathcal{P}_1(\mathbf{V}))|\forall \mathbf{A}_i,\mathbf{A}_j \in \alpha, \mathbf{A}_i \not\subseteq \mathbf{A}_j\}, \tag{6.47}$$

where $\mathcal{P}_1(\mathbf{V})$ denotes the power set of $\mathbf{V}$ without the empty set. Now it is possible to define a partial order on the set $\mathcal{A}(\mathbf{V})$ which reflects the structure of redundant information of the corresponding variables. A partially ordered set is a set with a relation that satisfies

reflexivity (each element is less than or equal to itself), antisymmetry (if an element is less than or equal to another and the converse is also true, then it follows that both are equal) and transitivity. The partial order $\preccurlyeq$ on $\mathcal{A}(\mathbf{V})$ is defined by the following relation

$$\forall \alpha, \beta \in \mathcal{A}(\mathbf{V}) : (\alpha \preccurlyeq \beta \iff \forall \mathbf{B} \in \beta \; \exists \mathbf{A} \in \alpha : \mathbf{A} \subseteq \mathbf{B}). \tag{6.48}$$

The set $\mathcal{A}(\mathbf{V})$ is also called an anti-chain or Sperner family on a set of $|\mathbf{V}| = n$ elements (Frank, 1980). The set $\mathcal{A}(\mathbf{V})$ together with the partial order $\preccurlyeq$ forms a lattice, which means that for any two elements of $\mathcal{A}(\mathbf{V})$ there is a unique least upper bound and a unique greatest lower bound in $\mathcal{A}(\mathbf{V})$ (Frank, 1980). Furthermore, any finite lattice is bounded, meaning there is an element that is less than every other element (bottom, $\bot$) and an element that is greater than every other element (top, $\top$). In the case of $\mathcal{A}(\mathbf{V})$ these are the the two sets $\bot = \{1\}\{2\}...\{n\}$ and $\top = \{12...n\}$. Returning to the redundancy measure the lattice gathers further meaning: the measure $I_\cap$ is monotonic with respect to the partial order, i.e. for any two elements $\alpha, \beta \in \mathcal{A}(\mathbf{V})$ where $\alpha \preccurlyeq \beta$ it follows that $I_\cap(Z;\alpha) \leq I_\cap(Z;\beta)$. This follows directly from the axioms for a measure of redundant information. As pointed out by Williams and Beer (2010) the lattice, also called *redundancy lattice*, already gives some insight into the structure of redundant information. The top element for example corresponds to the self-redundancy of $I(Z;X_\mathbf{V})$ and is thus an upper bound for $I_\cap$. The lattices for the cases $n = 2$ and $n = 3$ are illustrated in Figure 6.3.

6.4.1.2 *Partial Information Atoms*

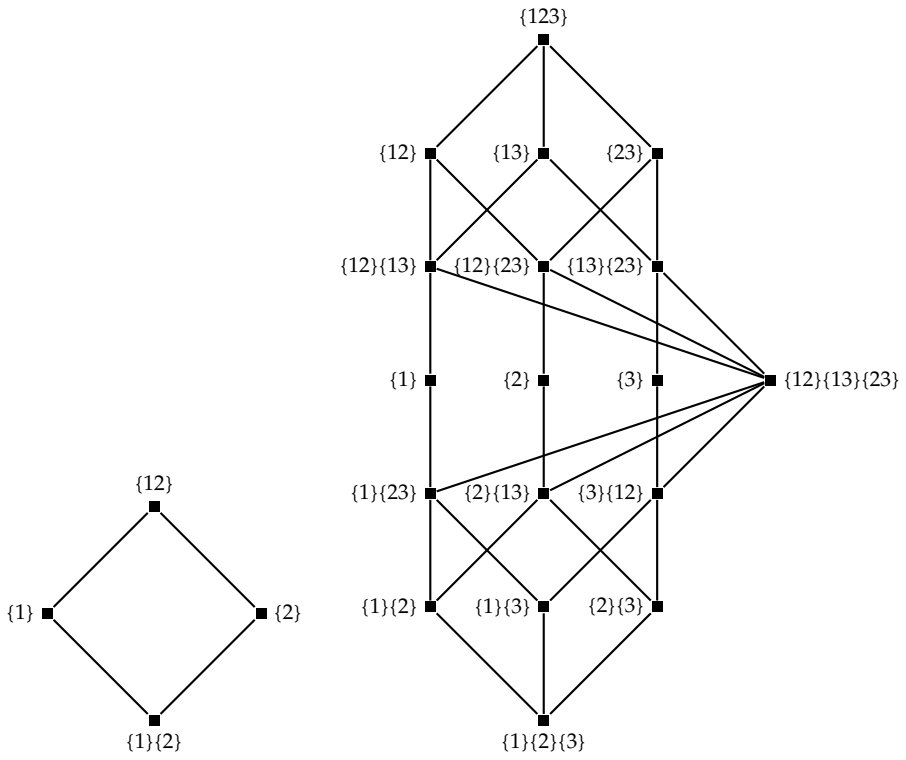
Starting from the bottom element of the redundancy lattice, which represents the amount of redundant information about Z that is contained in *all* individual variables X_i, going toward the top element, the associated amount of redundant information increases. It is now possible to ask the question what is the information that is contained redundantly in $\{2\}$ (i.e. the self-redundancy) but not redundantly in $\{1\}\{2\}$. This would be the aforementioned unique information of X_2. However, it is also possible to ask for the information redundantly contained in $\{1\}\{23\}$ but not redundantly in $\{1\}\{2\}$ or $\{1\}\{3\}$. These quantities are called partial information atoms, they are denoted by $\Pi^\cap_\mathbf{V}(Z;\beta)$ where $\beta \in \mathcal{A}(\mathbf{V})$. The definition depends on the underlying redundancy measure and is implicitly given by

$$I_\cap(Z;\alpha) = \sum_{\beta \preccurlyeq \alpha} \Pi^\cap_\mathbf{V}(Z;\beta). \tag{6.49}$$

Recursively the partial information atoms are now defined as

$$\Pi^\cap_\mathbf{V}(Z;\alpha) = I_\cap(Z;\alpha) - \sum_{\beta \prec \alpha} \Pi^\cap_\mathbf{V}(Z;\beta). \tag{6.50}$$

In the bivariate case, this leads to the decomposition of mutual information $I(Z;X_1,X_2)$ into four partial information atoms. Here the index contains only two elements, $\mathbf{V} = \{1,2\}$. Still following Williams and Beer (2010) the atomic terms are,



a) Redundancy lattice for $|V| = 2$

b) Redundancy lattice for $|V| = 3$

FIGURE 6.3 Redundancy lattices for different sizes of index sets. Vertices represent elements of $\mathcal{A}(\mathbf{V})$, edges are connected if an element is ``smaller'' with respect to the partial order $\preccurlyeq$ and there is no other element in $\mathcal{A}(\mathbf{V})$ that is smaller than the larger and larger than the smaller element.

- $\Pi^{\cap}_{\mathbf{V}}(Z;\{1\}\{2\}) = I_{\cap}(Z;X_1,X_2)$ which is the redundant information contained in X_1 and X_2 about Z,
- the unique information about Z which is only contained in X_1 or X_2 respectively, denoted as $\Pi^{\cap}_{\mathbf{V}}(Z;\{1\}) = I(Z;X_1) - I_{\cap}(Z;X_1,X_2)$ and $\Pi^{\cap}_{\mathbf{V}}(Z;\{2\}) = I(Z;X_2) - I_{\cap}(Z;X_1,X_2)$.
- and $\Pi^{\cap}_{\mathbf{V}}(Z;\{1,2\}) = I(Z;X_2,X_2) - I(Z;X_1) - I(Z;X_2) + I_{\cap}(Z;X_1,X_2)$, synergistic information, the information about Z that is only available if X_1 and X_2 are both known.

The sum of these terms is exactly the mutual information between Z and all sources, i.e.

$$I(Z;X_1,X_2) = \Pi^{\cap}_{\mathbf{V}}(Z;\{1\}\{2\}) + \Pi^{\cap}_{\mathbf{V}}(Z;\{1\}) + \Pi^{\cap}_{\mathbf{V}}(Z;\{2\}) + \Pi^{\cap}_{\mathbf{V}}(Z;\{1,2\}). \tag{6.51}$$

as well as

$$I(Z;X_1) = \Pi^{\cap}_{\mathbf{V}}(Z;\{1\}\{2\}) + \Pi^{\cap}_{\mathbf{V}}(Z;\{1\}) \tag{6.52}$$

and for X_2 respectively.

This decomposition allows a deep insight into the structure of multivariate information, however the decomposition is not necessarily non-negative for an arbitrary measure of redundancy. The decomposition has been introduced specifically to avoid the problems with interaction information. Therefore, I subscribe to the requirement that the redundancy measure should support a non-negative decomposition and negative partial information atoms should be avoided at all costs. In (Bertschinger et al., 2012) this is introduced as an additional axiom or property:

Local Non-negativity (LN) $\Pi_{\mathrm{V}}^{\cap} \geq 0$

For $I_{\min}$ the non-negativity of the corresponding PI-atom decomposition had been shown by Williams and Beer (2010) and was later discussed in greater detail by Williams (2011). Furthermore, they also introduce a closed form for the calculation of the partial information atoms based on $I_{\min}$. I will now continue to show that in the bivariate case the decomposition of the earlier introduced measure I_{red} also leads to a non-negative decomposition.

6.4.2 *Bivariate Decomposition Using Redundancy*

To show the non-negativity of the bivariate partial information decomposition using I_{red} the reader needs to be reminded that $I_{\mathrm{red}}(Z; X_1, X_2)$ is non-negative, as shown earlier. Furthermore, it follows from the self-redundancy and monotonicity axioms of the redundancy measure that $I_{\mathrm{red}}(Z; X_1, X_2) \leq I(Z; X_1)$ and with the same argument $I_{\mathrm{red}}(Z; X_1, X_2) \leq I(Z; X_2)$ which immediately implies that the unique information terms are non-negative. The following lemma now gives the non-negativity of the synergistic term:

Lemma 6.5. $I(Z; X_1, X_2) - I(Z; X_1) - I(Z; X_2) + I_Z^{\pi}(X_1 \searrow X_2) \geq 0.$

Proof. It is possible to reformulate the left hand side

$$
\begin{aligned}
& I(Z; X_1, X_2) - I(Z; X_1) - I(Z; X_2) + I_Z^{\pi}(X_1 \searrow X_2) && (6.53)\\
&= I(Z; X_1, X_2) - I(Z; X_2) \\
&\quad - \sum_x p(x_1) D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x \searrow X_2)}\right) && (6.54)\\
&= \sum_{x_1, x_2, z} p(z, x_1, x_2) \log \frac{p(z|x_1, x_2)}{p(z)} \\
&\quad - \sum_{x_1} \sum_{x_2, z} p(x_1|z, x_2) p(z, x_2) \log \frac{p(z|x_2)}{p(z)} \\
&\quad - \sum_{x_1} p(x_1) D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x \searrow X_2)}\right) && (6.55)\\
&= \sum_{x_1, x_2} p(x_1, x_2) \sum_z p(z|x_1, x_2) \log \frac{p(z|x_1, x_2)}{p(z)}
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{x_1,x_2} p(x_1,x_2) \sum_z p(z|x_1,x_2) \log \frac{p(z|x_2)}{p(z)} \\
&- \sum_{x_1} p(x_1) D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x \searrow X_2)}\right) && (6.56)\\
= &\sum_{x_1,x_2} p(x_1,x_2) D_{\mathrm{KL}}\left(P_{Z|x_1,x_2} \middle\| P_{Z|x_2}\right) \\
&- \sum_{x_1} p(x_1) D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x \searrow X_2)}\right) && (6.57)\\
= &\sum_{x_1} p(x_1) \left(\left(\sum_{x_2} p(x_2|x_1) D_{\mathrm{KL}}\left(P_{Z|x_1,x_2} \middle\| P_{Z|x_2}\right) \right) \right. \\
&\qquad \left. - D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right) \right) && (6.58)
\end{aligned}
$$

and now by the convexity of the Kullback-Leibler divergence:

$$
\begin{aligned}
\geq &\sum_{x_1} p(x_1) \left(D_{\mathrm{KL}}\left(\sum_{x_2} p(x_2|x_1) P_{Z|x_1,x_2} \middle\| \sum_{x_2} p(x_2|x_1) P_{Z|x_2} \right) \right. \\
&\qquad \left. - D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right) \right) && (6.59)\\
= &\sum_{x_1} p(x_1) \left(D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| R_{Z|x_1}\right) - D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right) \right) && (6.60)
\end{aligned}
$$

where $R_{Z|x_1} := \sum_{x_2} p(x_2|x_1) P_{Z|x_2} \in C_{\mathrm{cl}}(\langle X_2 \rangle_Z)$ and thus for all $x_1 \in \mathcal{X}_1$

$$
D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| R_{Z|x_1}\right) - D_{\mathrm{KL}}\left(P_{Z|x_1} \middle\| P_{Z|(x_1 \searrow X_2)}\right) \geq 0. \tag{6.61}
$$

□

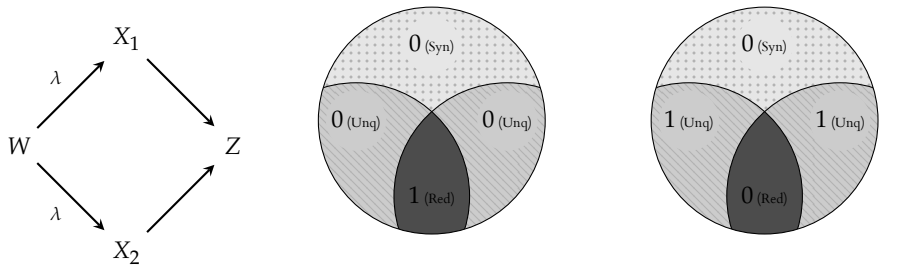
Thus the introduced measure can be used to decompose mutual information in a consistent manner.

6.4.3 *Examples*

I will now present examples of the partial information decomposition using I_{red} in the bivariate case. These examples will also serve as comparisons in the next section.

6.4.3.1 *Copying - From Redundancy to Uniqueness*

The first example is a very simple mechanism which simply copies the binary input variables X_1 and X_2 into Z, i.e. $Z = (X_1, X_2)$. However, I also add a control parameter $\lambda \in [0,1]$ which determines how correlated X_1 and X_2 are. This is done as follows: Let W be a uniformly distributed binary random variable, $p(x_1|w) = \frac{1}{2}\lambda + (1-\lambda)\delta_{x_1 w}$ and $p(x_2|w) = \frac{1}{2}\lambda + (1-\lambda)\delta_{x_2 w}$. The underlying model is the Bayesian network as depicted in Figure 6.4a. For $\lambda = 1$, X_1 and X_2 are independent, as the Bayesian network describes



a) Bayesian model b) PI-diagram for $\lambda = 0$ (RDN) c) PI-diagram for $\lambda = 1$ (UNQ)

FIGURE 6.4 Copy Example. Complete redundancy and complete uniqueness using I_{red}.

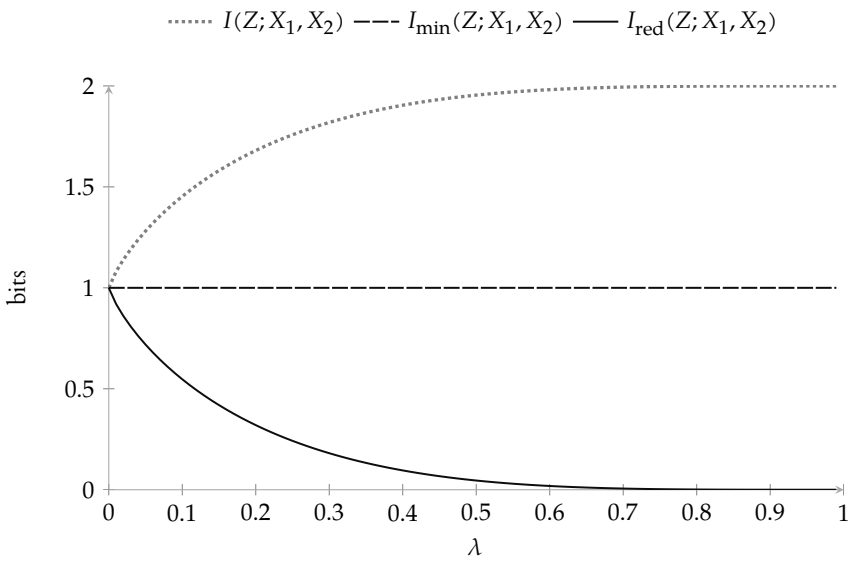
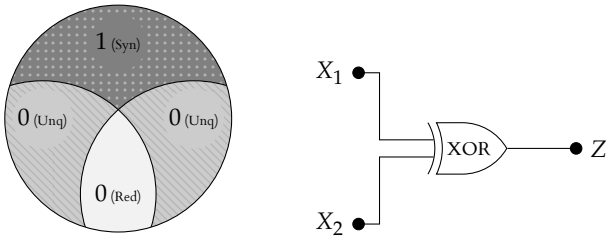


FIGURE 6.5 Comparison of total mutual information $I(Z; X_1, X_2)$ (dotted gray line), the new redundancy measure I_{red} (solid line) and $I_{\min}$ (dashed line) for varying values of λ, where λ controls the correllation between X_1 and X_2. It can be seen $I_{\min}$ measures a constant amount of redundancy and therefore does not distinguish between redundancy and uniqueness with varying λ as desired, whereas I_{red} does.

the complete model, recovering the example 'UNQ (Unique Information)' as introduced by Griffith (2011). At the other extreme, where $\lambda = 0$, X_1 and X_2 are identical copies of W and therefore Z is equivalent to W from an information-theoretic point of view. This is also reflected in the decomposition as in this case $I(Z; X_1, X_2) = I(W; X_1, X_2)$ and $I_{\text{red}}(Z; X_1, X_2) = I_{\text{red}}(W; X_1, X_2)$. This is the example 'RDN (Redundant Information)' from (Griffith, 2011). By varying λ the entropy of the outcome Z is varied and at the same time unique information is exchanged for redundancy. Figure 6.4bc illustrates the decomposition at both extremal values of λ and it can be seen that the resulting values of I_{red} coincide with the proposed values in (Griffith, 2011). The effect of changing λ is shown in Figure 6.5.



a) PI-diagram b) circuit diagram

FIGURE 6.6 XOR Example. A purely synergistic mechanism.

6.4.3.2 *XOR*

The XOR gate ($\oplus$), is a classical example for the appearance of synergy, in the sense of the whole being more than the sum of the individuals. Thus it is expected to only observe synergistic information, as the result is only known if both inputs are available, and the uncertainty given one input is the same as knowing no input at all. Here, the inputs are uniformly distributed independent binary random variables X, Y and the output is $Z = X \oplus Y$. In fact, in this case $I_{\text{red}}(Z; X, Y) = I_{\min}(Z; X, Y) = 0$ resulting in the purely synergistic decomposition as illustrated in Figure 6.6. The redundancy measure vanishes here because $P_Z = P_{Z|x} = P_{Z|(x \searrow Y)}$, as well as $P_Z = P_{Z|y} = P_{Z|(y \searrow X)}$, i.e. the information about the outcome of Z is zero even if one input is known. This would change if correlation between X and Y is introduced. Note that I_{red} defines the redundancy, other terms are all derived by the decomposition.

6.4.3.3 *AND - Mechanisms at Work*

The next example is the AND gate, $Z = X_1 \wedge X_2$. This turns out to be an interesting case, because it demonstrates the subtle difference between redundant information that is due to the 'ignorance' of the mechanism with respect to the source, and redundancy that is already apparent in the sources. In (Griffith, 2011 and Griffith and Koch, 2012) it is argued that vanishing mutual information between the sources X_1 and X_2 themselves implies vanishing redundant information[1]. This feature is also shared by the synergy measure introduced in (Griffith and Koch, 2012). However, here I would like to embrace a different view on redundant information: even if the sources are independent, there can be a correlation in the change of the distribution over Z given observations in X_1 and X_2 respectively. Observing one input does not give any information about the other input, but part of the information gain about the distribution of the output can be the same as one gets from the other input alone. In particular in the case of the AND gate, observing a 0 in either input

1 "However, because X_1 and X_2 are independent, [...], thus necessitating there is zero redundant information [...].",(Griffith, 2011)

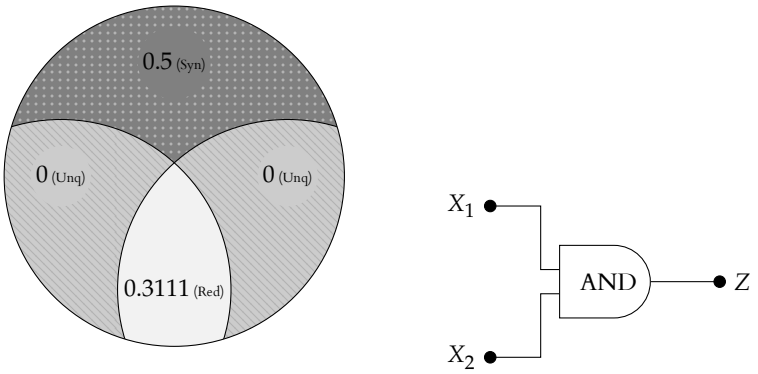


FIGURE 6.7 AND Example. The total mutual information is $I(Z; X_1, X_2) = 0.811278$.

leads to $p(z = 0) = 1$. As a result of calculating the redundancy for this example one obtains $I_{\text{red}}(Z; X_1, X_2) = I_{\min}(Z; X_1, X_2) = 0.311278$. Figure 6.7 illustrates the decomposition of the total mutual information for this example.

I will denote redundant information that is only due to the mechanism, as it is the case with the redundancy in the AND gate, *mechanistic redundancy*. Contrary to this I will call redundant information that already appears in the inputs *source redundancy*. Redundancy in the source must already manifest itself in the mutual information between the inputs. At this point I do not give a rigorous definition for these terms, as it can be seen in the next example, there are cases where it is not clear how to separate both. However, if there is positive redundant information $I_{\text{red}} > 0$ but vanishing mutual information between the sources, all redundant information can be attributed to mechanistic redundancy.

6.4.3.4 *Summing Dice*

Consider an example where two dice are thrown (cubic dice, with numbered sides from 0 to 5), represented by the random variables D_1, D_2. The results are summed and the dice D_1 and D_2 are uniformly distributed and independent. There are several ways to sum the results: simply add the two results — this would lead to results ranging from 0 to 10 where 5 is the most probable result and 0 or 10 the least probable result — or multiply the result of the first die by 6 to get a uniform distribution of all numbers ranging from 0 to 35. Indeed, I will look at all intermediate summations here, defined by $R = \alpha D_1 + D_2$ where $\alpha \in \{1, 2, 3, 4, 5, 6\}$. The hypothesis motivating this example was that for the direct summation ($\alpha = 1$) there is a positive amount of redundancy between D_1 and D_2 with respect to R, because knowing the roll of one die gives 'overlapping' information (in the same direction within the space of distributions) with the roll of the other die about the final result. The redundancy should then decrease if α is increased, up to the point where

$\alpha = 6$ and the sum of both dice rolls is isomorphic to the joint variable of the two dice rolls, i.e. $6D_1 + D_2 \simeq (D_1, D_2)$. Indeed, this is reflected in the redundancy $I_{\text{red}}(R; D_1, D_2)$. In Figure 6.8 an additional parameter λ was added, that controls how correlated the two dice are, in the same way as λ was introduced in the copy example in Section 6.4.3.1 to control the correlation between the input variables. For $\lambda = 1$ they are independent and it can be seen that the redundancy increases with decreasing α, on the other extreme $\lambda = 0$ the dice are completely correlated. In this case the redundancy that is already existent in the source ($I(D_1, D_2) \approx 2.58$) shadows all redundancy otherwise induced through the mechanism and hence there is no difference in the redundancy value for different values of α.

6.5 COMPARISONS

After the construction and presentation of the proposed measure of redundant information, I will now continue by comparing it to the measures introduced earlier on (see Section 6.2) starting with the measure of minimal information I_{min}. The reader is reminded that the construction above only covers the bivariate case and therefore the comparison also is mainly focused on the bivariate case. An outlook towards a generalization of the ideas used for the construction of the bivariate measure follows in Section 6.8.

6.5.1 *Relation to Minimal Information*

The development of the redundancy measure I_{red} was motivated by the shortcomings of I_{min} outlined earlier in the chapter. However, the construction still tries to capture the same idea of redundancy and thus it is no surprise that there are some cases where I_{red} and I_{min} coincide. In general there is a tendency of I_{min} to overestimate redundancy and it seems that I_{min} is an upper bound for I_{red} in most cases. There are a few exceptions, but

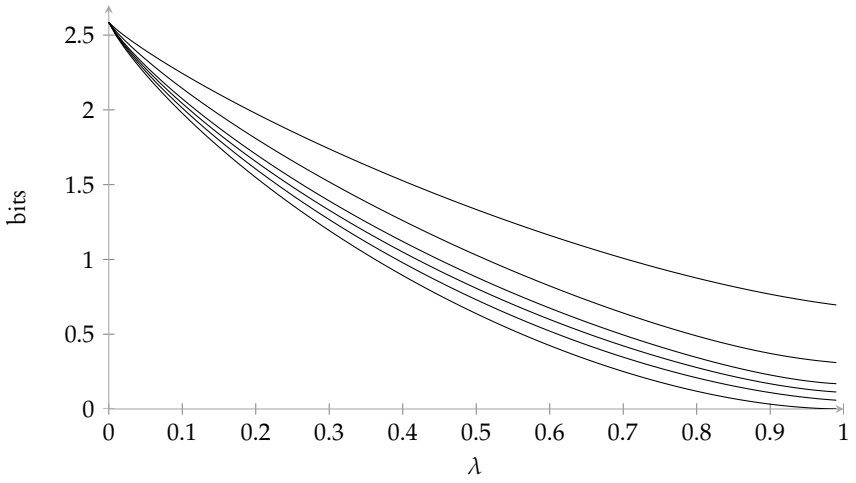


FIGURE 6.8 Plot of the redundant information $I_{\text{red}}(R; D_1, D_2)$ depending on the correlation λ between the two dice D_1 and D_2. From top to bottom the summation coefficient is $\alpha = 1, \ldots, 6$. It can be seen that for independent dice $\lambda = 1$ the amount of redundancy depends on the mechanism that is used to sum the results, whereas on the other extreme, all redundancy comes from the correlation of the sources.

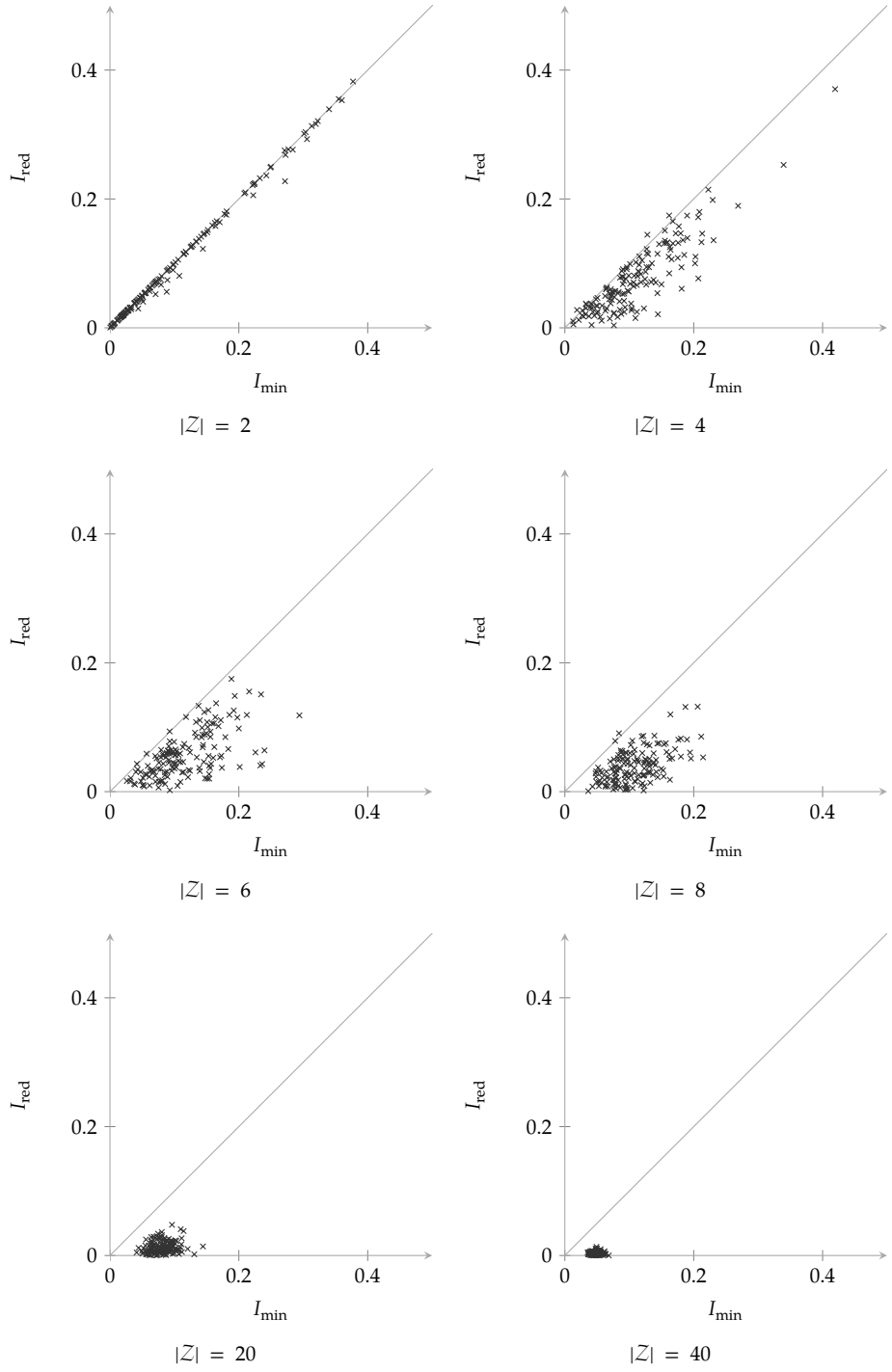


FIGURE 6.9 Comparison of I_{min} and I_{red} for randomly drawn distributions $p(x, y, z)$ with $|\mathcal{X}| = |\mathcal{Y}| = 3$ fixed sized sets, plotted for different sizes of $\mathcal{Z}$. The change of $|\mathcal{Z}|$ also changes the dimension of the simplex where the distributions P_Z are contained in. Note that as the dimension of $\mathcal{Z}$ goes up, I_{min} gets larger in comparison to I_{red}. The distributions were drawn using a uniform distribution on a random subsimplex of $\Delta(X, Y, Z)$. The subsimplex was selected in each draw randomly with the probability of $p(x, y, z) = 0$ being 0.5 for each triple (x, y, z).

they are due to numerical instabilities. The overestimation of redundancy by $I_{\min}$ becomes predominant if the dimension of Z is increased (see Figure 6.9). The explanation for this is that, the higher the dimension of the space gets, the larger the error becomes which results from not taking directionality into account.

6.5.2 *Axioms Revisited*

In Section 6.2.5 additional desired properties of redundancy measures were introduced. Together all properties (or axioms) are

Weak Symmetry (S_0) $I_\cap$ is symmetric with respect to permutations the A_i's.

Self-Redundancy (I) $I_\cap(Z;A) = I(Z;X_A)$.

Monotonicity (M) $I_\cap(Z;A_1,...,A_{k-1},A_k) \leq I_\cap(Z;A_1,...,A_{k-1})$ with equality if $A_{k-1} \subseteq A_k$.

Non-negativity (N) $I_\cap \geq 0$

Identity (Id_2) $I_\cap\left((X_{A_1},X_{A_2});A_1,A_2\right) = I(X_{A_1};X_{A_2})$

Strong Symmetry (S_1) $I_\cap$ is symmetric with respect to permutations of Z and the A_i's.

Left Monotonicity (LM) $I_\cap(Z;A_1,...,A_{k-1},A_k) \leq I_\cap(Z,Z';A_1,...,A_{k-1},A_k)$.

Left Chain Rule (LC) $I_\cap(Z,Z';A_1,...,A_{k-1},A_k) = I_\cap(Z;A_1,...,A_{k-1},A_k) + I_\cap(Z';A_1,...,A_{k-1},A_k|Z)$.

Local Non-negativity (LN) $\Pi^\cap_V \geq 0$

Some of these properties are implying each other. For example left monotonicity **(LM)** follows from the left chain rule **(LC)** and identity **(Id_2)** from the left chain rule **(LC)** and local non-negativity **(LN)**. Furthermore there is the following theorem:

Theorem 6.1. (Bertschinger, Rauch, Olbrich and Jost) *There is no measure of shared information [redundant information] that satisfies* **(S_1)**, **(M)**, **(I)** *and* **(LN)**.

This is a quite interesting result as it shows that strong symmetry **(S_1)** is incompatible with the features that the PI-decomposition using $I_{\min}$ and I_{red} offers. Strong symmetry is also conflicting with the mechanistic perspective established above, i.e. redundancy that is induced by the ignorance of the mechanism from which source information is coming, which is clearly not symmetric under a permutation of A_i with Z. However, a strongly symmetric measure could provide what is needed to quantify source redundant information, i.e. information about Z that is redundantly available in all sources (and not by means of

any mechanism). Consequently, this also means that the measure of redundant information derived from synergistic mutual information as introduced by Griffith and Koch (2012) cannot be used for a PI-decomposition as laid out by Williams and Beer (2010), because it fulfils **(S_1)**, **(M)** and **(I)** .

6.5.3 *Measures of Shared Information*

In Section 6.2.5 I postponed the definition of the two measures of shared information introduced in (Bertschinger et al., 2012). I will now give the definitions of these measures. Both measure depend on a family of probability distributions on Z, denoted by $\{P_{x_{A_1}|\ldots|x_{A_k}}\}_{x_{A_1},\ldots,x_{A_k}}$ and defined as

$$P_{x_{A_1}|\ldots|x_{A_k}} = \arg\min \left\{ D_{\mathrm{KL}} \left(\sum_i \lambda_i P_{Z|x_{A_i}} \middle\| P_Z \right) \middle| \lambda_i > 0, \sum_i \lambda_i = 1 \right\}. \tag{6.62}$$

Geometrically speaking, for each event in the source variables $X_{A_1}, \ldots, X_{A_k}$, $P_{x_{A_1}|\ldots|x_{A_k}}$ is the closest distribution to P_Z in the convex closure of $\{P_{Z|x_{A_1}}, \ldots, P_{Z|x_{A_k}}\}$ and the construction is almost dual to the construction of the projections $P_{Z|(x_1 \searrow X_2)}$ in the bivariate case. The idea is to capture the least informative distribution about Z for every possible outcome of the input variables. Now the two measure are defined as follows:

$$I_{\mathrm{slg}}(Z; A_1, \ldots, A_k) := \sum_{x_{A_1},\ldots,x_{A_k},z} p(z, x_{A_1}, \ldots, x_{A_k}) \log \frac{p_{x_{A_1}|\ldots|x_{A_k}}(z)}{p(z)}, \tag{6.63}$$

$$I_{\mathrm{skl}}(Z; A_1, \ldots, A_k) := \sum_{x_{A_1},\ldots,x_{A_k}} p(x_{A_1}, \ldots, x_{A_k}) D_{\mathrm{KL}} \left(P_{x_{A_1}|\ldots|x_{A_k}} \middle\| P_Z \right). \tag{6.64}$$

However, both measures have shortcomings, I_{slg} violates monotonicity and I_{skl} can lead to negative synergy in the PI-decomposition. Therefore, they cannot be used for a PI-decomposition.

6.5.4 *Left Monotonicity*

That I_{red} violates strong symmetry is conceptually justified as explained above (and in greater detail in the next section). But I_{red} also violates left monotonicity **(LM)**. This is problematic because left monotonicity appears to be a very desirable property: Extending the output variable (e.g. Z becomes Z, Z' where the marginal distribution on Z stays fixed) should not decrease the amount of redundancy as the original output is changed in no way. The example below shows a calculation where the extension of Z to Z, Z' reduces the redundancy between the input variables X_1 and X_2. At the moment I suspect that there are geometric effects due to the increase of the dimension of the space $\Delta(Z)$, which are not accounted for by the construction. However, I did not succeed in constructing a measure that fullfils **(S_0)**, **(I)**, **(M)**, **(LN)** and **(LM)**. This means that the search for measures of redundancy and synergy is still a very open field with room for improvements. In the

FIGURE 6.10 Conditional distributions visualized in $\Delta(Z)$ on the unit interval.

meantime I_{red} serves as a good candidate, especially if in the specific application the ouput variable Z has a fixed dimension and even more so if the mechanism $P_{Z|x_1,x_2}$ does not change between comparisons using I_{red}.

In the following example introduced by Bertschinger et al. (2012) the left monotonicity is violated by I_{red}. The joint probability distribution for this example is given in Table 6.1. It is easy to calculate the corresponding redundancy values $I_{\text{red}}(Z; X_1, X_2)$,$I_{\text{red}}(Z, Z'; X_1, X_2)$.

X_1	X_2	Z	Z'	p
0	0	0	0	1/6
0	1	0	0	1/6
0	1	0	1	1/6
1	1	0	1	1/6
1	0	1	1	2/6

TABLE 6.1 Joint distribution of an example where $I_{\text{red}}(Z; X_1, X_2) > I_{\text{red}}(Z, Z'; X_1, X_2)$, thus violating left monotonicity **(LM)**.

In this example it is easily possible in an analytical way, the space of distributions over Z is one-dimensional ($\Delta(Z)$ is isomorphic to the unit interval), thus P_Z, $P_{Z|x_1}$ and $P_{Z|x_2}$ can be represented as points on the unit interval for each x_1 and x_2 respectively. In this specific case $P_Z \simeq 2/3$ $P_{Z|x_1=0} \simeq 1$, $P_{Z|x_1=1} \simeq 1/3$, $p(Z|x_2 = 0) \simeq 1/3$ and $P_{Z|x_2=1} \simeq 1$, if the distribution $p(z = 0) = 1$, $p(z = 1) = 0$ is mapped to 1 on the unit interval. A visualization of the one dimensional case with only Z as output can be found in Figure 6.10. From this it is clear that the mutual projections are actually just identity mappings (i.e. $P_{Z|(x_1 \searrow X_2)} = P_{Z|x_1}, \ldots$) and hence the marginal and conditional probabilities can simply be calculated from Table 6.1,

$$I_{\text{red}}(Z; X_1, X_2) = I_Z^{\pi}(X_1 \searrow X_2) = I_Z^{\pi}(X_2 \searrow X_1) \tag{6.65}$$

$$= I(Z; X_1) = I(Z; X_2) \tag{6.66}$$

$$= \frac{1}{2}\log_2\frac{3}{2} + \frac{1}{6}\log_2\frac{1}{2} + \frac{1}{3} \approx 0.4591. \tag{6.67}$$

To calculate $I_{\text{red}}(Z, Z'; X_1, X_2)$ it is easier to calculate the projections $P_{Z,Z'|(x_1 \searrow X_2)}$ and $P_{Z,Z'|(x_1 \searrow X_2)}$ numerically. This results in

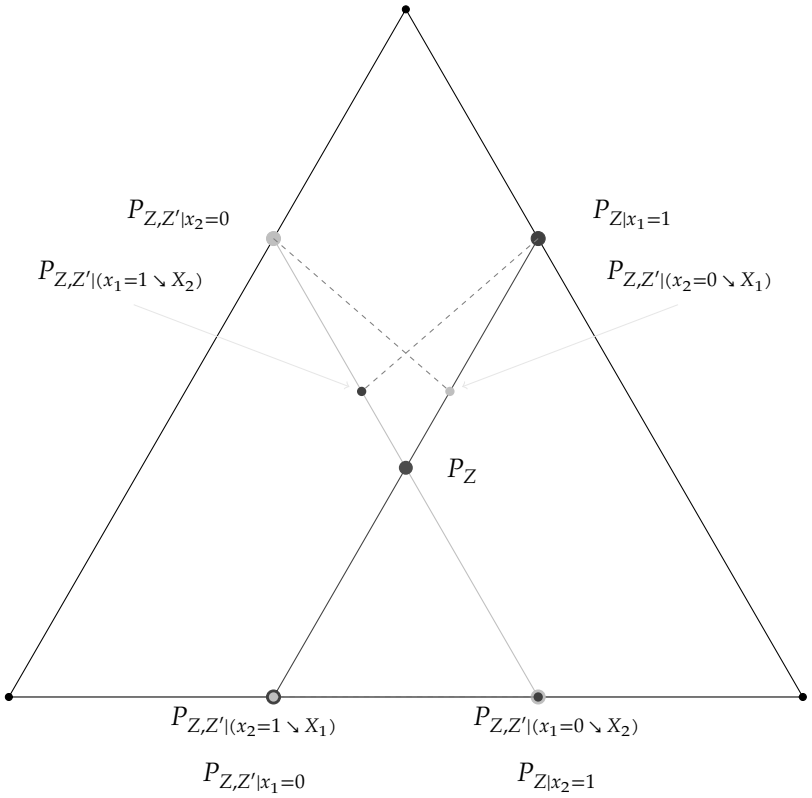


FIGURE 6.11 Illustration of the construction of projective information for binary input variables. The illustration shows why left monotonicity does not hold for I_{red}.

$$P_{Z,Z'|(x_1=0 \searrow X_2)} = P_{Z,Z'|x_2=1}, \tag{6.68}$$

$$P_{Z,Z'|(x_1=1 \searrow X_2)} = \frac{2}{3} P_{Z,Z'|x_2=0} + \frac{1}{3} P_{Z,Z'|x_2=1}, \tag{6.69}$$

$$P_{Z,Z'|(x_2=0 \searrow X_2)} = \frac{1}{3} P_{Z,Z'|x_1=0} + \frac{2}{3} P_{Z,Z'|x_1=1}, \tag{6.70}$$

$$P_{Z,Z'|(x_2=1 \searrow X_2)} = P_{Z,Z'|x_1=0}, \tag{6.71}$$

which are visualized in Figure 6.11. Thus the value for the redundant information is now (the equality of both projective information terms is only a result of the explicit example, not a general property)

$$I_{\text{red}}(Z, Z'; X_1, X_2) = I^{\pi}_{Z,Z'}(X_1 \searrow X_2) = I^{\pi}_{Z,Z'}(X_2 \searrow X_1) \tag{6.72}$$

$$= \frac{1}{6} \log_2 \frac{2}{3} + \frac{1}{6} + \frac{1}{3} \log_2 \frac{4}{3} \approx 0.2075. \tag{6.73}$$

And hence $I_{\text{red}}(Z; X_1, X_2) > I_{\text{red}}(Z, Z'; X_1, X_2)$ which shows that I_{red} does not fulfill left monotonicity.

6.6 MECHANISTIC & SOURCE REDUNDANCY

In Section 6.4.3.4 it was shown that the measure I_{red} also captures redundancy that is due to the mechanism of the particular system that is being examined. This is also the reason why the measure violates strong symmetry **(S_1)**. A mechanism is not necessarily symmetric with respect to the permutation of sources and output, however the redundancy with respect to the output variables, that is already existent in the sources should be strongly symmetric as it ignores the mechanism in terms of redundancy. A candidate measure for source redundancy is the redundancy measure induced by the synergistic mutual information (Griffith and Koch, 2012) as introduced in Section 6.2.4:

$$I_{\text{src}}(Z; X_1, X_2) = I(Z; X_1) + I(Z; X_2) + S(Z; X_1, X_2) - I(Z; X_1, X_2) \tag{6.74}$$

$$= I(Z; X_1) + I(Z; X_2) - I_{\cup}(Z; X_1, X_2). \tag{6.75}$$

To quantify the amount of redundant information that is induced by the mechanism it does not suffice to simply subtract the source redundancy I_{src} from the overall amount of redundancy I_{red}. For example in the case of the AND gate (see Section 6.4.3.3), assume that the two inputs are identical copies of each other and thus completely correlated. In this case there the mechanism still induces redundant information, but due to the correlation of the sources this does not increase the overall amount of redundant information. The already apparent redundant information in the inputs is 'shadowed' by the redundancy in the mechanism. To measure the amount of redundancy induced by the mechanism I propose the following measure

$$I_{\text{mec}}(Z; X_1, X_2) = I_{\text{red}}(\hat{Z}; \hat{X}_1, \hat{X}_2) \tag{6.76}$$

where $p(\hat{z}|\hat{x}_1, \hat{x}_2) = p(z|x_1, x_2)$ and $p(\hat{x}_1, \hat{x}_2) = p(x_1)p(x_2)$, thus separating the source and removing any redundancy about Z available in the sources. This is an intervention at the source introducing independent sources. There are a few desired properties these measures should fulfil.

Conjecture 6.1. *$I_{\text{src}} \leq I_{\text{red}}$, $I_{\text{mec}} \leq I_{\text{red}}$ and $I_{\text{src}} + I_{\text{mec}} \geq I_{\text{red}}$.*

In the current form with $I_{\text{src}} = I(Z; X_1) + I(Z; X_2) - I_{\cup}(Z; X_1, X_2)$ these are only conjectures. Numerical experiments did not give conclusive results because the optimization procedure used in the calculation of I_{src} remained in a local minimum for all the tests I ran. Nonetheless, any proposed measures should fulfil these inequalities: The first two inequalities state that source and mechanistic redundancy cannot exceed the total amount of redundant information, which reflects the idea that redundant information measured by either of the two measures is also measured by I_{red}. Assuming that there is no third source for redundancy, after all, all there is are the sources and the mechanism, the third inequality states that redundant information is either already existing in the sources or induced by the mechanism, though possibly combined.

A last remark on the distinction between source and mechanistic redundancy: The difference $(I_{\text{src}} + I_{\text{mec}}) - I_{\text{red}}$ is the amount of information that is 'shadowed' by the mechanism. This quantity is very interesting as it is linked to the problem of overdetermination in causal information flow measures. Ay and Polani (2008) and Janzing et al. (2012) introduce measures of causal information flow and both use a similar technique in the construction of the measures as done for the mechanistic redundancy above. Both articles use interventions by setting distributions of certain nodes in the causal Bayesian network to their marginals. However, in the example of the AND gate, each measure would measure a causal influence from both sources, but give no further information about the structure of the causal information flow. Using redundancy measures would already allow to distinguish unique causes from redundant causes, but being able to quantify the amount of 'shadowed' redundancy might lead to a quantification of overdetermination that is not caused by a correlation of the causes.

6.7 INFORMATION TRANSFER

In the following section I will show that I_{red} shares all the properties that I_{min} possesses with respect to the decomposition of *transfer entropy*. In (Williams and Beer, 2011) the partial information decomposition is used to introduce new measures of information transfer. The measures are based on a decomposition of *transfer entropy*. Transfer entropy, introduced by Schreiber (Schreiber, 2000), is defined for two random processes X_t and Y_t as

$$T_{Y \to X} = I(X_{t+1}; Y_t | X_t). \tag{6.77}$$

It measures the influence of the process Y at time t on the state of the process X in the next time step. One can also take a longer history of Y_t and X_t into account instead. The right hand side, known as *conditional mutual information*, is

$$I(X_{t+1}; Y_t | X_t) = I(X_{t+1}; Y_t, X_t) - I(X_{t+1}; X_t). \tag{6.78}$$

As the conditional entropy is the difference of two mutual information terms, the PI-decomposition can be used to decompose each mutual information term. Hence by the vanishing of PI-atoms, the transfer entropy can be decomposed into two non-negative components. The decomposition is illustrated in Figure 6.12. I will use a slightly different notation here, instead of using an index set, as there are only two variables, the variable names are used directly. Let $\mathbf{V} = \{X_t, Y_t\}$ then it follows from (6.51) and (6.52) that

$$T_{Y \to X} = \Pi'_{\mathbf{V}}(X_{t+1}; \{Y_t\}) + \Pi'_{\mathbf{V}}(X_{t+1}; \{X_t, Y_t\}). \tag{6.79}$$

The first term denotes all information that uniquely comes from Y_t, called *State Independent Transfer Entropy* (SITE) by Williams and Beer (2011). The second term on the other hand denotes information that comes from Y_t but depends on the state of X_t and thus is called *State Dependent Transfer Entropy* (SDTE) in (Williams and Beer, 2011). I will now apply both

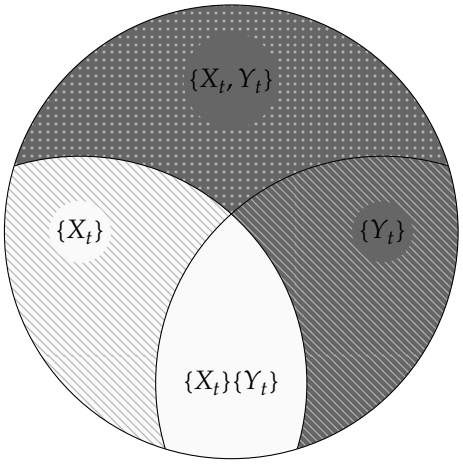


FIGURE 6.12 PI-diagram for the decomposition of transfer entropy into PI-atoms. The coloured areas denote the transfer entropy.

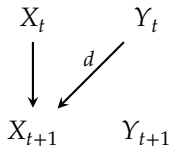


FIGURE 6.13 Bayesian network of the first example process. If $x_t = 0$ then x_{t+1} is a copy of y_t, if $x_t = 1$ then the bit of x_{t+1} is a flipped copy y_t. The probability that the bit is flipped in the copy is denoted by d.

measures $I_{\min}$ (with corresponding PI-atoms $\Pi_{\mathbf{V}}$) and I_{red} (with corresponding PI-atoms $\Pi'_{\mathbf{V}}$) as the underlying redundancy measure for the decomposition and compare the results.

6.7.1 *Coupled Markov Processes Examples*

The following two examples are used to show the difference of the decomposition when using I_{red} instead of $I_{\min}$. The first one revisits an example from (Williams and Beer, 2011) where X and Y are two binary, coupled Markov random processes. The process Y is uniformly i.i.d. and $x_{t+1} = y_t$ if $x_t = 0$, moreover

$$p(x_{t+1} = y_t | x_t = 1) = 1 - d, \tag{6.80}$$

$$p(x_{t+1} = 1 - y_t | x_t = 1) = d. \tag{6.81}$$

So $d \in [0, 1]$ controls whether there is any dependence on the previous state of X. If d vanishes X is simply a copy of Y, see Figure 6.13 for a Bayesian network of the process. In this case the redundancy between Y_t and X_t with respect to X_{t+1} also vanishes as X_t contains no information about X_{t+1}, but at the same time $I(X_{t+1}; X_t, Y_t) = I(X_{t+1}; Y_t)$ so the synergy

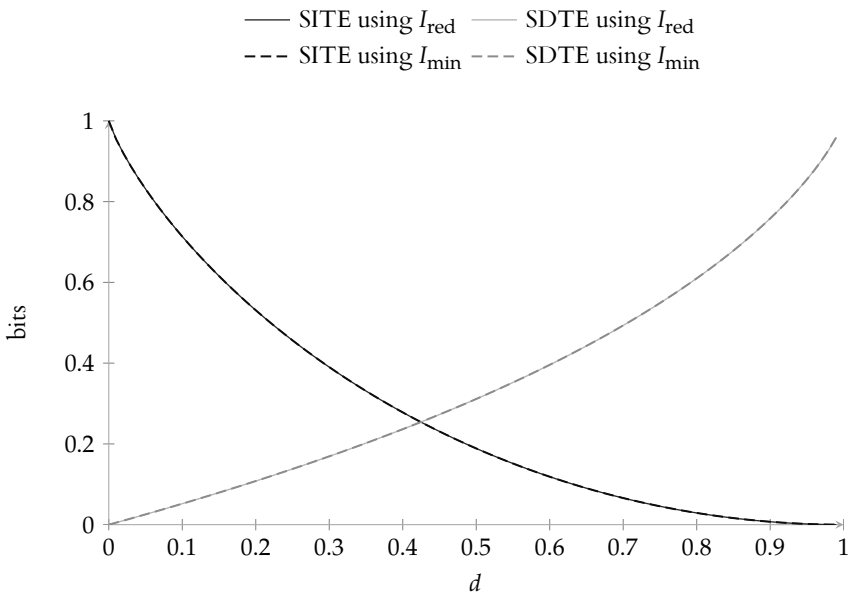


FIGURE 6.14 Decomposition of transfer entropy $T_{Y\to X}$ for the first example process. The plot shows SITE (solid black line using I_{red}, dashed black line using $I_{\min}$) and SDTE (solid gray line using I_{red}, dashed gray line using $I_{\min}$) given d. It can be seen that both decompositions coincide for this process.

also vanished and thus the example shows only state-independent transfer entropy. Increasing d now reduces the overall mutual information $I(X_{t+1}; X_t, Y_t)$ but the information that Y_t contains about X_{t+1} is decreasing at a faster rate, while the redundancy stays constantly zero with varying d. The state independent transfer entropy $\Pi'_{\mathbf{V}}(X_{t+1}; \{Y_t\})$ in this example is equal to $I(X_{t+1}; Y_t)$ and thus decreases while the state dependent transfer entropy (synergy) $\Pi'_{\mathbf{V}}(X_{t+1}; \{X_t, Y_t\})$, here the difference $I(X_{t+1}; X_t, Y_t) - I(X_{t+1}; Y_t)$, increases with increasing d (compare with Figure 6.14). This also explains why the decompositions of transfer entropy using either measure (I_{red}, $I_{\min}$) coincide, the redundancy is constantly zero and the change of the PI-atom is only driven by the change of mutual information terms.

The second example, constructed for this specific purpose, is more intricate. First of all it shows the difference between the two measures, but it is also a good example of the subtlety of redundancy in mechanisms. Consider the following combined process (X_t, Y_t) and the process Z_t where Z_t for all t are uniformly i.i.d. random variables, X_{t+1} is a copy of X_t and

$$p(y_{t+1}|y_t, z_t) = (1-d)\delta_{y_t y_{t+1}} + d\delta_{z_t y_{t+1}}. \tag{6.82}$$

The process Y, copies with probability d the value of Z_t and with probability $(1-d)$ the value of Y_t to Y_{t+1}. Now the transfer entropy $T_{Z\to(X,Y)}$ is measured, see Figure 6.15 for a Bayesian network of the process.

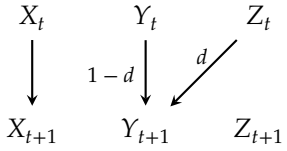


FIGURE 6.15 Bayesian network of the second example process. X_t is a parallel and independent process, the only information transfer between the processes is from Z_t to Y_{t+1}.

6.7.1.1 *Comparison of Transfer Entropy of both Examples*

It can be seen in Figure 6.16 that the two decompositions coincide for $d \leq 0.5$. For $d = 0$ the two processes are completely independent which is reflected in the vanishing overall transfer entropy in this case. On the other extreme using $d = 1$, the decomposition using I_{red} gives complete state-independent transfer entropy while the decomposition using $I_{\min}$ sees total state-dependent transfer entropy. In this case the decompositions disagree completely and I argue here that my new measure reflects the process much better. With $d = 1$ the process always copies Z_t to Y_{t+1}, which is completely independent of (X_t, Y_t). Specifically, $I_{\min}$ mistakenly measures redundancy between X_t and Z_t with regard to $(X, Y)_{t+1}$. Following the definition of synergy and (6.52) this is then reflected in the vanishing state-independent transfer entropy for all d (larger redundancy means more synergy and less unique information, given that the mutual information stays constant).

The fact that $I_{\min}$ measures more redundancy has the same reason why $I_{\min}$ measures redundancy between independent X and Y with respect to $Z = (X, Y)$, namely it compares changes in different directions in the space of distributions. The parallel and independent process X_t lets $I_{\min}$ see a dependency between the two processes X_t and Z_t that I argue does not exist. Considering the transfer entropy $T_{Z \to Y}$ from Z_t to Y_t only, ignoring the process X_t completely, it can be seen in Figure 6.17 that the decomposition now coincides with the decomposition of $T_{Z \to (X,Y)}$ using I_{red} (solid lines in Figure 6.16).

Nonetheless, this does not yet explain the quite unusual non-differentiable shape of the state-independent transfer entropy, which only is positive for $d > 0.5$. This is surprising because up to $d = 0.5$ all transfer entropy is considered to be state-dependent, even though with probability d the state of Y_{t+1} takes on the state of Z_t. As the process X_t was only used to demonstrate that using $I_{\min}$ for the decomposition measures state dependencies in the transfer-entropy that are not there, I will now leave X_t aside and only consider the process (Y_t, Z_t) as described above.

To understand the shape of the graph of state-dependent transfer entropy of this process, it is helpful to have a look at the mutual information $I(Y_{t+1}; Z_t)$ (dotted gray line in Figure 6.18) and the redundancy $I_{\text{red}}(Y_{t+1}; Y_t, Z_t)$ (solid black line in Figure 6.18). From (6.52) it follows that the state-independent transfer entropy (solid black line in Figure 6.16 and dashed black line in Figure 6.17) is now the difference of these two terms (compare with Figure 6.12).

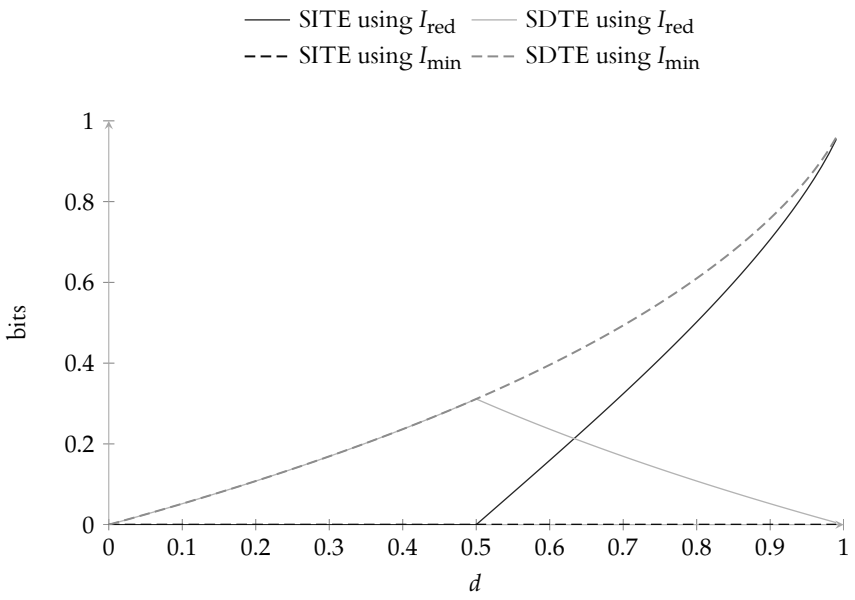


FIGURE 6.16 Decomposition of transfer entropy $T_{Z\to(X,Y)}$ for the second example process. The plot shows SITE (solid black line using I_{red}, dashed black line using I_{min}) and SDTE (solid gray line using I_{red}, dashed gray line using I_{min}).

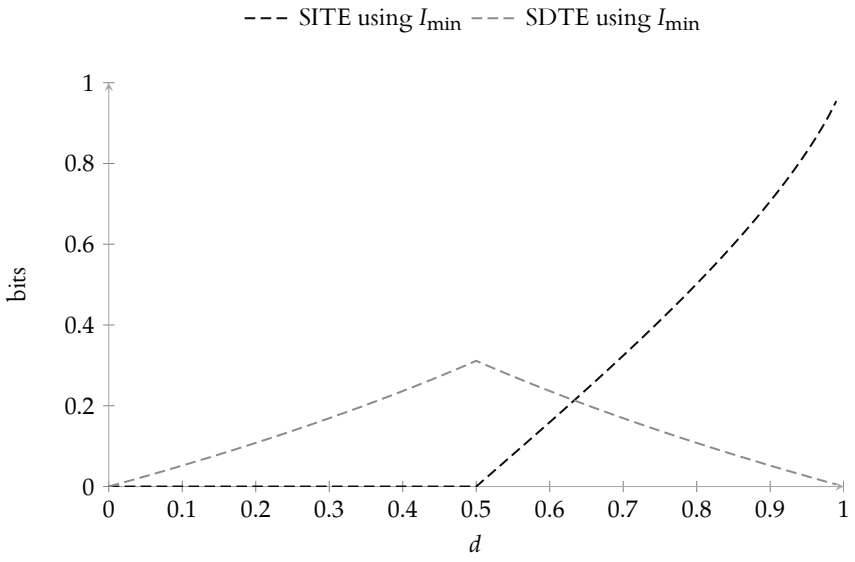


FIGURE 6.17 Decomposition of transfer entropy $T_{Z\to Y}$ for the second example process. The plot shows SITE (dashed black line using I_{min}), SDTE dashed gray line using I_{min}).

The increase of mutual information $I(Y_{t+1}; Z_t)$ is obvious from the definition of the process. For $d = 0$ both processes are independent and for $d = 1$ it follows that $Y_{t+1} = Z_t$. It is also clear that the redundant information with respect to Y_{t+1} needs to be zero at the extremal points $d \in \{0, 1\}$, because at these points the value of Y_{t+1} depends either on Y_t $(d = 0)$ or Z_t $(d = 1)$ and therefore either $I(Y_{t+1}; Z_t) = 0$ or $I(Y_{t+1}; Y_t) = 0$ both of which are upper bounds for the redundancy.

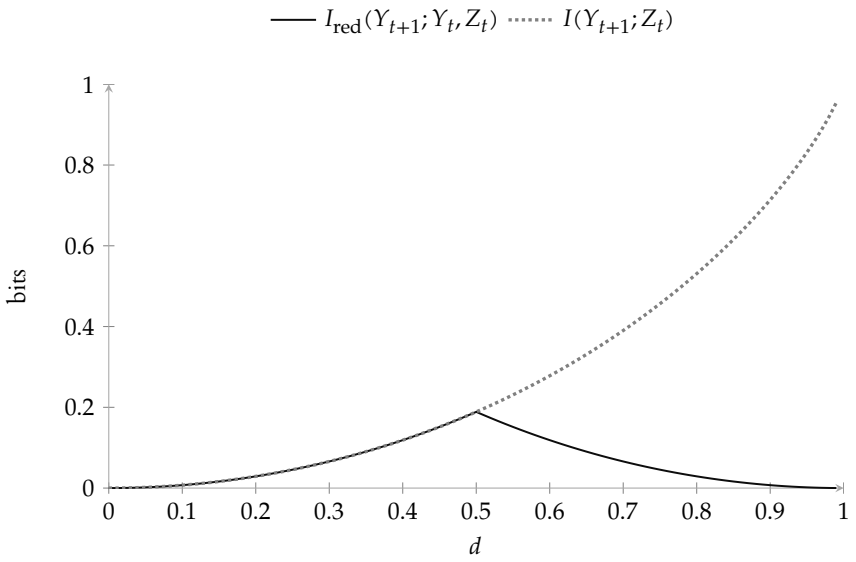


FIGURE 6.18 The plot shows $I(Y_{t+1}; Z_t)$ (dotted gray line) and $I_{\text{red}}(Y_{t+1}; Y_t, Z_t)$ (solid black line) for the second example process.

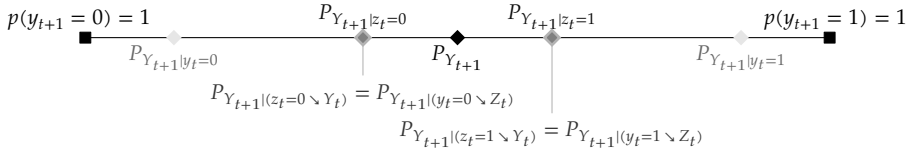
On the other hand for $d = 0.5$ the state of either process at time t tells something about the distribution of Y_{t+1} and because the space of distributions of Y_{t+1} is one-dimensional, this must be information about a change in the same direction, so there is positive redundancy. Observing one of the outcomes necessarily contributes to some extent to the prediction of the outcome of Y_{t+1}. Now it is possible to show this more rigourously, in particular

$$p(y_{t+1}|y_t) = \frac{d}{2}\delta_{y_{t+1}(1-y_t)} + \left(1 - \frac{d}{2}\right)\delta_{y_{t+1}y_t}\,, \tag{6.83}$$

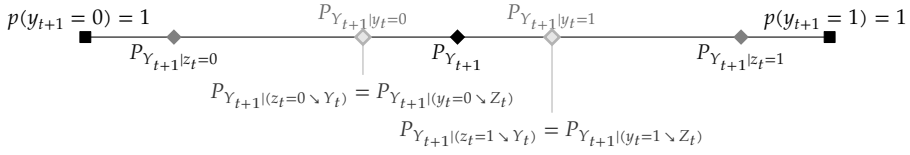
$$p(y_{t+1}|z_t) = \frac{1-d}{2}\delta_{y_{t+1}(1-z_t)} + \frac{1+d}{2}\delta_{y_{t+1}z_t}\,. \tag{6.84}$$

are the conditional distributions given the current state of either Y_t or Z_t. To calculate $I_{\text{red}}(Y_{t+1}; Y_t, Z_t)$ projected information $I^{\pi}_{Y_{t+1}}(Z_t \searrow Y_t)$ needs to be calculated, as well as $I^{\pi}_{Y_{t+1}}(Y_t \searrow Z_t)$ because the redundancy is the minimum of both terms. As the space of distributions $\Delta(Y_{t+1})$ is one dimensional (it is simply the unit interval), a simple illustrative argument can be made to compute $P_{Y_{t+1}|(z_t=0 \searrow Y_t)}$, $P_{Y_{t+1}|(z_t=1 \searrow Y_t)}$, $P_{Y_{t+1}|(y_t=0 \searrow Z_t)}$ and $P_{Y_{t+1}|(y_t=1 \searrow Z_t)}$, which are the terms that are needed to calculate projected information. From the illustration in Figure 6.19 it can be seen that for $d \leq 0.5$, $P_{Y_{t+1}|(z_t=0 \searrow Y_t)} = P_{Y_{t+1}|(y_t=0 \searrow Z_t)} = P_{Y_{t+1}|z_t=0}$ and $P_{Y_{t+1}|(z_t=1 \searrow Y_t)} = P_{Y_{t+1}|(y_t=1 \searrow Z_t)} = P_{Y_{t+1}|z_t=1}$. Inserted into (6.14) it follows that $I^{\pi}_{Y_{t+1}}(Z_t \searrow Y_t) = I^{\pi}_{Y_{t+1}}(Y_t \searrow Z_t) = I(Y_{t+1}; Z_t)$ for $d \leq 0.5$. This explains why there is no SITE for $d \leq 0.5$, as the SITE is the difference between the redundancy $I_{\text{red}}(Y_{t+1}; Y_t, Z_t)$ and $I(Y_{t+1}; Z_t)$.

Conversely $I^{\pi}_{Y_{t+1}}(Z_t \searrow Y_t) = I^{\pi}_{Y_{t+1}}(Y_t \searrow Z_t) = I(Y_{t+1}; Y_t)$ for $d \leq 0.5$. As $I(Y_{t+1}; Z_t)$ and $I(Y_{t+1}; Y_t)$ are perfectly symmetric, which explains the form of the redundant information as in (solid black line in Figure 6.18). Thus, even though Z_t and Y_t are completely



a) $\Delta(Y_{t+1})$ for $d \leq 0.5$



b) $\Delta(Y_{t+1})$ for $d \geq 0.5$

FIGURE 6.19 Illustration of the conditional distributions of Y_{t+1} for the second example process in the two cases $d \leq 0.5$ and $d \geq 0.5$. The line represents the one dimensional simplex, i.e. the space of probability distributions over Y_{t+1} denoted by $\Delta(Y_{t+1})$ where Y_{t+1} is a binary valued random variable. The black diamond represents the marginal distribution of $p(y_{t+1})$ and the shaded diamonds the conditionals given specific values of Y_t and Z_t. It can now be seen that the projections are always equal to the conditional distributions closer to the marginal of Y_{t+1}. In particular, the projections are the same, no matter in which direction the projection is done (from Y_t to Z_t or vice versa).

independent, the mechanism, as discussed earlier in Section 6.6, creates redundancy with respect to Y_{t+1}.

6.7.2 *Control Theory and Redundancy*

Ashby (1956) proposed and Touchette and Lloyd (2000) confirmed that there is a natural link between control theory and information theory. As shown by Touchette and Lloyd (2004), for a process, with initial state X and final state X', and a controller C which are linked by the probability distribution $p(x'|x,c)$, the conditional mutual information $I(X';C|X)$ (which is the transfer entropy from the controller to the system) is a measure of controllability. Williams and Beer show in (Williams and Beer, 2011) that the decomposition of transfer entropy using $I_{\min}$ as a redundancy measure has a close relation to the notion of open-loop controllability. I will now show, that this is still the case if I_{red} is used to decompose transfer entropy.

Perfect controllability, as defined in (Touchette and Lloyd, 2004), means that for all initial states $x \in \mathcal{X}$ and final states $x' \in \mathcal{X}$ there exists a control state $c \in \mathcal{C}$ such that $p(x'|x,c) = 1$. The following equivalence is then shown in (Williams and Beer, 2011)

Lemma 6.6. *A system is perfectly controllable iff for any x' there exists a distribution $p(c|x)$ such that $p(x') = 1$ for any distribution $p(x)$.*

It follows also that if a system is perfectly controllable, there exists an x' such that $p(x'|x) = 1$ for each $x \in \mathcal{X}$, see (Williams and Beer, 2011) for a proof. Now, a system has perfect open-loop controllability iff it has perfect controllability and $I(X;C) = 0$. Moreover, in (Williams and Beer, 2011) it is shown that the following theorem holds:

Theorem 6.2. (Williams and Beer) *A system is perfectly open-loop controllable iff it is perfectly controllable with vanishing state-dependent transfer entropy (using $I_{\min}$) from C to X'.*

Furthermore, this theorem still holds in the case where the decomposition using I_{red} is used. To prove the theorem the following lemma is needed. It is shown in (Williams and Beer, 2011) that the condition of the lemma is fulfilled for any perfect open-loop controller and thus proves the direct part of the theorem (perfect open-loop controllability implies perfect controllability with zero SDTE using I_{red} as a redundancy measure):

Lemma 6.7. *If*

$$\forall x' \in \mathcal{X} \forall x \in \mathcal{X} \forall c \in \mathcal{C} : p(x'|x,c) = p(x'|c) \tag{6.85}$$

then the SDTE from C to X' is zero.

Proof. From the definition of the partial information decomposition it follows that

$$\begin{aligned} \Pi'(X';\{C,X\}) = I(X';X,C) - I(X';X) \\ - I(X';C) + \Pi'(X';\{C\},\{X\}). \end{aligned} \tag{6.86}$$

Using the definition of the redundancy measure it follows that

$$\begin{aligned} \Pi'(X';\{C,X\}) \leq I(X';X,C) - I(X';X) \\ - I(X';C) + I^{\pi}_{X'}(X \searrow C). \end{aligned} \tag{6.87}$$

The synergy is non-negative and now the right hand side can be reformulated as in (6.58). But with $p(x'|x,c) = p(x'|c)\ \forall x, x' \in \mathcal{X}, c \in \mathcal{C}$ the positive Kullback-Leibler divergences in (6.58) all vanish. Therefore $\Pi'(X';\{C,X\}) = 0$. □

For the converse direction, perfect controllability and vanishing SDTE (from C to X') imply perfect open-loop controllability, the following lemma needs to be proved:

Lemma 6.8. *If a system is perfectly controllable with a distribution $p(c|x)$ then $I_{\text{red}}(X';X,C) = 0$.*

Proof. From Lemma 6.6 it follows that $p(x') = 1$ for some $x' \in \mathcal{X}$ as well as $p(x'|x) = 1$ for all $x \in \mathcal{X}$ and therefore $C_{\text{cl}}(\langle X \rangle_Z)$ in $\Delta(X')$ is just $\{P_{X'}\}$ which implies $I^{\pi}_{X'}(C \searrow X) = 0$. Thus it follows that $I_{\text{red}}(X';X,C) = 0$. □

Thus, for the converse direction, starting with perfect controllability and vanishing SDTE, the following equality holds

$$0 = \Pi'(X'; \{C, X\}) \tag{6.88}$$
$$= I(X'; X, C) - I(X'; X) \tag{6.89}$$
$$- I(X'; C) + I_{\text{red}}(X'; X, C) \tag{6.90}$$
$$= I(X'; X, C) - I(X'; X) - I(X'; C) \tag{6.91}$$
$$= \sum_{x,c,x'} p(x', x, c) \log \frac{p(x'|x, c)p(x')}{p(x'|c)p(x'|x)}, \tag{6.92}$$

as $p(x'|x) = p(x')$ because of perfect controllability,

$$= \sum_{x,c,x'} p(x', x, c) \log \frac{p(x'|x, c)}{p(x'|c)} = I(X; X'|C). \tag{6.93}$$

The definition of perfect controllability in (Touchette and Lloyd, 2004) gives that for each $x \in \mathcal{X}$ and $x' \in \mathcal{X}$ there exists a $c \in \mathcal{C}$ such that $p(x'|x, c) = 1$. From Eq. (6.93) it follows now that for any $x' \in \mathcal{X}$ there exists a $c \in \mathcal{C}$ such that $p(x'|c) = 1$, as otherwise the logarithm in all summands with $p(x, x', c) > 0$ would not vanish in Eq. (6.93). It is shown in (Williams and Beer, 2011) that the proposition that there exists a $c \in \mathcal{C}$ such that $p(x'|c) = 1$ for any $x' \in \mathcal{X}$ is equivalent to open-loop controllability. Together with what was shown above, this shows that Theorem 6.2 also holds if I_{red} is the underlying redundancy measure and the relation between open-loop controllability and decomposition of transfer entropy is transferable to the newly constructed measure.

6.8 MULTIVARIATE EXTENSIONS

So far everything in this chapter, except Section 6.4, was about bivariate measures of redundancy or bivariate applications of multivariate measures. The construction of I_{red} is inherently bivariate, so the question for a multivariate extension of the measure arises naturally. A naive approach to a multivariate extension would be simply taking the minimum of pairwise redundancies, i.e.

$$I_{\text{red}}(Z; A_1, \ldots, A_k) := \min_{i,j} I_{\text{red}}(Z; A_i, A_j). \tag{6.94}$$

And indeed, this construction fulfils the **(S_0)**, **(M)**, **(I)** axioms and even **(LN)** which allows a non-negative decomposition of multivariate mutual information into partial information atoms, as I_{min} does. However, a similar problem as with the bivariate case of I_{min} appears. Consider the three mutually independent random variables X_1, X_2, X_3. Now, with the minimum construction, the redundancy $I_{\text{red}}(X_1, X_2, X_3; \{1, 2\}, \{2, 3\}, \{3, 1\})$ would be

$$\min\{I(X_1, X_2; X_2, X_3), I(X_2, X_3; X_3, X_1), I(X_3, X_1; X_1, X_2)\} \tag{6.95}$$

which in turn is

$$\min\{H(X_1), H(X_2), H(X_3)\}. \tag{6.96}$$

It can be seen that the identity axiom **($\mathbf{Id_2}$)**, even though it gives an upper bound for redundancy (by means of the monotonicity axiom **(M)**) still is not enough to ensure that a multivariate redundancy measure fully captures the concept of redundancy. Namely that it would be desirable to have

$$I_{\text{red}}(X_1, X_2, X_3; \{1,2\}, \{2,3\}, \{3,1\}) = 0 \tag{6.97}$$

here.

To achieve this I will extend the concept of projected information. For this each conditional distribution in $\Delta(Z)$ will be projected onto the intersection of projections of convex closures. In agreement with the notation used for the bivariate measure, for any $Q \in \Delta(Z)$ and any subset $B \subseteq \Delta(Z)$, $P_{Z|(Q \searrow B)}$ denotes the projection of Q onto the convex closure of B in $\Delta(Z)$ (this is consistent with the notation $P_{Z|(x \searrow Y)}$ where Y denotes the set of distributions $\{P_{Z|y} \in \Delta(Z) \,|y \in \mathcal{Y}\}$ and x is used to represent the distribution $P_{Z|x} \in \Delta(Z)$). In the same way the projected information can now be defined for any subset $B \subseteq \Delta(Z)$ as

$$I_Z^{\pi}(X \searrow B) := \sum_{z,x} p(z,x) \log \frac{p_{Z|(x \searrow B)}(z)}{p(z)}. \tag{6.98}$$

The projection of the convex closure of $B_1 \subseteq \Delta(Z)$ onto the convex closure of $B_2 \subset \Delta(Z)$ will now be defined as

$$B_1 \searrow B_2 := C_{\text{cl}}\left(\bigcup_{Q \in C_{\text{cl}}(B_1)} \{P_{Z|(Q \searrow B_2)}\} \right). \tag{6.99}$$

The proposed definition of multivariate redundancy is now as follows

$$I_{\text{red}}(Z; A_1, \ldots A_k) := \min_{1 \leq i \neq j \leq k} I_Z^{\pi}\left(X_{A_i} \searrow \left(\bigcap_{l=1}^{k} (X_{A_l} \searrow X_{A_j}) \right) \right). \tag{6.100}$$

I will show that this is equivalent to the following definition

$$I_{\text{red}}(Z; A_1, \ldots A_k) = \min_{1 \leq i \neq j \leq k} I_Z^{\pi}\left(X_{A_i} \searrow \left(\bigcap_{l=1, l \neq j}^{k} (X_{A_l} \searrow X_{A_j}) \right) \right) \tag{6.101}$$

and for $k \geq 3$ it is even possible to leave one more term out of the intersection:

$$I_{\text{red}}(Z; A_1, \ldots A_k) = \min_{1 \leq i \neq j \leq k} I_Z^{\pi}\left(X_{A_i} \searrow \left(\bigcap_{l=1, l \neq j, l \neq i}^{k} (X_{A_l} \searrow X_{A_j}) \right) \right). \tag{6.102}$$

The first equivalence is true because $X_{A_j} \searrow X_{A_j} = C_{\text{cl}}(X_{A_j})$ and

$$\bigcap_{l=1, l \neq j}^{k} (X_{A_l} \searrow X_{A_j}) \subseteq C_{\text{cl}}(X_{A_j}). \tag{6.103}$$

To show the second equivalence, let

$$B = \bigcap_{l=1, l\neq j, l\neq i}^{k} (X_{A_l} \searrow X_{A_j}). \tag{6.104}$$

By the same argument as for the first equivalence $B \subseteq C_{\mathrm{cl}}(X_{A_j})$. Now

$$P_{Z|(x_{A_i} \searrow B)} \in C_{\mathrm{cl}}(B) \subseteq C_{\mathrm{cl}}(X_{A_j}), \tag{6.105}$$

and by definition

$$X_{A_i} \searrow X_{A_j} := C_{\mathrm{cl}}\left(\bigcup_{Q \in C_{\mathrm{cl}}(X_{A_i})} \{P_{Z|(Q \searrow X_{A_j})}\} \right). \tag{6.106}$$

As $P_{Z|x_{A_i}} \in C_{\mathrm{cl}}(X_{A_i})$ it follows that also $P_{Z|(x_{A_i} \searrow B)} \in X_{A_i} \searrow X_{A_j}$.
Hence,

$$I_Z^{\pi}\left(X_{A_i} \searrow B\right) = I_Z^{\pi}\left(X_{A_i} \searrow B \cap (X_{A_i} \searrow X_{A_j})\right) \tag{6.107}$$

and the equivalence of the redundancy definitions follows.

As in the bivariate case self redundancy is simply defined as $I_{\mathrm{red}}(Z;A) = I_Z^{\pi}(X_A \searrow X_A)$. Furthermore, the bivariate version of the multivariate measure coincides with the earlier introduced bivariate redundancy measure. For $k = 2$

$$\begin{aligned} I_{\mathrm{red}}(Z;A_1,A_2) &= \min\{I_Z^{\pi}\left(X_{A_1} \searrow X_{A_1} \searrow X_{A_2}\right), I_Z^{\pi}\left(X_{A_2} \searrow X_{A_2} \searrow X_{A_1}\right)\} && (6.108)\\ &= \min\{I_Z^{\pi}\left(X_{A_1} \searrow X_{A_2}\right), I_Z^{\pi}\left(X_{A_2} \searrow X_{A_1}\right)\} && (6.109) \end{aligned}$$

as $X_{A_1} \searrow X_{A_2}$ contains all possible projections of distributions in $C_{\mathrm{cl}}(X_{A_1})$ onto $C_{\mathrm{cl}}(X_{A_2})$ (correspondingly for $X_{A_2} \searrow X_{A_1}$). The definition is obviously weakly symmetric **(S_0)** and the inequality of the monotonicity **(M)** can be easily shown by observing that

$$\bigcap_{l=1}^{k} (X_{A_l} \searrow X_{A_j}) \subseteq \bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_j}) \tag{6.110}$$

for all $j \in \{1, \ldots, k-1\}$, i.e. there are only terms less than or equal to existing projected information terms in the minimization. To show equality if $A_{k-1} \subseteq A_k$ I will show that all the terms over which the minimization is performed for $I_{\mathrm{red}}(Z;A_1,\ldots,A_k)$ are greater than or equal to a term over which the minimization is performed for $I_{\mathrm{red}}(Z;A_1,\ldots,A_{k-1})$. The first case to look at, is the case where $j = k$ (and thus $i \neq k$ by definition). Here

$$\begin{aligned} & C_{\mathrm{cl}}(X_{A_{k-1}}) \subseteq C_{\mathrm{cl}}(X_{A_k}) && (6.111)\\ \Rightarrow\ & X_{A_l} \searrow X_{A_{k-1}} \subseteq X_{A_l} \searrow X_{A_k} \text{ for all } l \in \{1, \ldots, k-1\} && (6.112)\\ \Rightarrow\ & \bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_{k-1}}) \subseteq \bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_k}) = \bigcap_{l=1}^{k} (X_{A_l} \searrow X_{A_k}) && (6.113)\\ \Rightarrow\ & I_Z^{\pi}\left(X_{A_i} \searrow \left(\bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_{k-1}})\right)\right) \leq I_Z^{\pi}\left(X_{A_i} \searrow \left(\bigcap_{l=1}^{k} (X_{A_l} \searrow X_{A_k})\right)\right). && (6.114) \end{aligned}$$

The next case is $j \neq k$ and $i \neq k$: If it is possible to show that

$$X_{A_{k-1}} \searrow X_{A_j} \subseteq X_{A_k} \searrow X_{A_j} \tag{6.115}$$

then

$$\bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_j}) = \bigcap_{l=1}^{k} (X_{A_l} \searrow X_{A_j}) \tag{6.116}$$

and the equality of all terms where $j \neq k$ and $i \neq k$ would follow. Now, notice that $X_{A_{k-1}} \searrow X_{A_j} \subseteq X_{A_k} \searrow X_{A_j}$ immediately follows from the definition of the projection operator ($\searrow$) as $C_{\mathrm{cl}}(X_{A_{k-1}}) \subseteq C_{\mathrm{cl}}(X_{A_k})$ in this case. The case $j \neq k$ and $i = k$ follows from Lemma 6.3 on page 123. Using the representation $X_{A_k} = (X_{A_{k-1}}, W)$ for some random variable W. Lemma 6.3 gives

$$I_Z^\pi \left(X_{A_{k-1}} \searrow \left(\bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_j}) \right) \right) \leq I_Z^\pi \left((X_{A_{k-1}}, W) \searrow \left(\bigcap_{l=1}^{k-1} (X_{A_l} \searrow X_{A_j}) \right) \right) \tag{6.117}$$

for all $j \neq k$ and gives the desired result which can be plugged into the definition of the multivariate measure.

Hence, it is in theory possible to extend the bivariate measure to a multivariate measure. However, this has some drawbacks. The projection of a convex closure onto another convex closure is not easily analytically expressed, even in this specific case, i.e. it is not a convex polytope anymore, though it is still a convex set. Thus, projecting onto the intersection of several of these projected sets makes the optimization problem much more complex to solve. Furthermore, I was not able to show local non-negativity **(LN)** for the partial information decomposition using the multivariate definition of I_{red}.

6.9 DISCUSSION

The motivation for this chapter was to overcome the shortcomings of current measures of redundancy and synergy, which are quantities of strong interest in relation to self-organization and information theory. I introduced a new measure for bivariate redundant information. Redundant information between two random variables is information that is shared between two variables with respect to a third variable. The measure is conceptually motivated by measuring similarities in the 'direction of change' in the outcome distribution, depending on which input is observed. I proved that the construction adheres to properties of redundancy as stated in the literature, and can be used for a non-negative decomposition of mutual information. The measure is closely related to the concept of *minimal information* as introduced in (Williams and Beer, 2010).

I demonstrated in several examples that I_{red} follows several intuitions about redundancy. Furthermore, it is possible to decompose *transfer entropy* as considered in (Williams and Beer, 2011); in particular I showed that using *minimal information* instead of *redundant information*

to decompose *transfer entropy* can lead to the detection of seemingly state-dependent transfer entropy which contradicts intuition. I was also able to prove that the results about open-loop controllability from (Williams and Beer, 2011) are also applicable to the decomposition using I_{red}. Thus the measure is able to serve as a replacement for the bivariate version of minimal information.

A particular insight of the new definition is the emphasis on mechanisms in the concept of redundant information, which has been rather neglected in the literature so far. I linked bivariate redundant information in the case of a copying mechanism to the mutual information between the input variables. Thus, I identified redundant information that already appears in the inputs with *source redundancy*, contrary to redundant information that is only due to the mechanism, as demonstrated in the AND-gate or the readout process with the decomposition of transfer entropy. I refer to the latter kind of redundancy as *mechanistic redundancy*. This is in contrast to the redundancy measure proposed in (Griffith and Koch, 2012) which does not capture such *mechanistic redundancy*. I proposed initial steps towards the separation of mechanistic and source redundant information, which includes the problem of 'shadowed' redundancy with possible applications for the measure of causal information flows.

A practical limitation that currently exists is the restriction to a bivariate measure. In general, however, there are applications where it is interesting to be able to compute redundant information between more than two variables (Williams and Beer, 2010 and Flecker et al., 2011). However, the geometric structure for this problem gets significantly more complex, so that my proposed multivariate extension is not much more than a theoretical exercise at the moment. Together with the violation of left monotonicity **(LM)** by I_{red} and I_{min} this highlights that the case of redundant information is not closed yet. There are still opportunities to find improvements in this area, yet this chapter presents a tool for the investigation of bivariate redundancy in a multitude of scenarios.

In the introduction to this chapter, I explained that I started to study the embodiment of agents in the perception-action loop using redundancy, but quickly became aware of the limitations of currently available measures. Although my work resulted in a novel approach to measure redundancy, its application to the embodiment of agents within the perception-action loop did not lead to profoundly new insights. The problem one is facing here, is that the world state usually encodes also the state of the agent and thus it is only possible to differentiate between internal cognitive information processing and 'other' information processing, which is performed by either the environment or the embodiment (one could call this embodied open-loop control and captures the same notion as the term morphological computation does). Nonetheless, I am certain that there are a lot of relevant applications of redundancy and synergy measures to the field of self-organization. For example, the redundancy lattice might be used to devise optimal information sharing networks in multi-agent systems or help to quantify the robustness of complex systems.

Initial findings (Lizier et al., 2013 and Flecker et al., 2011) lead me to believe that the study of self-organizing systems from a redundancy/synergy perspective only has started.

CONCLUSION

7

» *A conclusion is the place where you got tired of thinking.* «
MARTIN H. FISCHER, Washingtonian

In this thesis, I set out to develop an information-theoretic framework for the investigation of the self-organization of agent collectives with a particular focus on morphogenetic processes, that is, the spatial self-organization of a collective towards shapes. Here, I first want to summarize what was done and discuss the results in a broader setting than I did in the individual chapters. Moreover, an overview of what is left to be done is given.

7.1 SUMMARY

Chapter 3 introduced different concepts of self-organization from which I chose observer based self-organization to work with. The choice was based on a literature review of available quantitative definitions of self-organization. The outlook that the measure could be easily adopted to the continuous domain as well as its grounding in information-theoretic concepts were crucial factors in this choice. Then, I continued with a comparison of several estimation methods for multi-information which is used for the quantification of observer self-organization. In this comparison, methods based on the binning of continuous data performed very poorly, while kernel based estimators and the KSG method gave better results. However, only the KSG estimator was able to give reliable results in high dimension settings ($d \geq 20$), where the kernel based estimator report information between uncorrelated data.

In Chapter 4, I introduce a model of particle dynamics, similar to the one used in (Doursat, 2008b), where each particle in the collective can be of a type and particles interact with each other depending on their types. These dynamics are shown to result in the particle collective forming in various shapes, depending on the parameters of the particle interactions. Using alignment procedures and observations about invariant properties of the dynamics, I show that multi-information of the particle locations can be estimated from a comparatively small amount of samples. This, in turn, gives a way to quantify the observer self-organization expressed by the particle collective. I used this, to study the effects of the number of particles, number of types and cut-off radii on the observable self-organization. In these experiments it was possible to see that the amount self-organization of the collective depends on the number of types (for a particle collective of a fixed number of particles) and the cut-off radius that limits interactions between particles. A large ratio of types to particles, usually means that interactions are very inhomogeneous leading to a low amount of correlation between the particles. On the other extreme, the organization is also comparatively low if there is only a single type. For collective with particles of a single type the organization also depends strongly on the size of the collective, as for several sizes differently shaped equilibrium configurations lead to a lower correlation among the individual particles. It was also possible to observe a dependence between the amount of self-organization and

the cut-off radius. For a small cut-off radius the self-organization decreased, however for a small ratio of types to particles this effect was not as strong as for collectives with a large ratio of types to particles.

In Chapter 5, I presented the relevant information formalism developed by Polani et al. (2006) and extended it to the multi-agent setting. I showed how shape formation can be formulated in terms of a reward function in the relevant information and MDP setting. Using the information-theoretic formulation of control theoretic principles (Touchette and Lloyd, 2004,2000), I was able to derive a relation between self-organization of agent collectives and the information processed in each agent exists. In the second part of Chapter 5 I looked into the coordination of agent actuators and sensors and the implications of shared control for the relevant information formalism and presented initial findings for small collectives of two agents.

Chapter 6 switched the focus towards measures of redundant information. I showed from examples that earlier approaches did not capture some intuitive requirements about redundancy, which lead to the formulation of an identity axiom for measures of bivariate redundant information. I then pursued by developing a bivariate measure of redundant information, that fulfils the axioms first stated by Williams (2011), as well as the additional identity axiom. The measure is based on the notion of information projections and I was able to show that it complies with the partial information decomposition by Williams and Beer (2010). Furthermore, I proved that the decomposition of transfer entropy, using the here introduced measure, results in the same properties regarding controllability as the measure of minimal information by Williams and Beer (2010). The last part of Chapter 6 is concerned with the construction of a multivariate measure of redundant information based on information projections. I proposed a construction of a measure, that fulfils the axioms, which however suffers from the lack of an easy way to compute actual numerical results. There is no proof available yet, that would show that it leads to a positive partial information decomposition.

7.2 DISCUSSION

7.2.1 *Self-Organization*

Observer based self-organization is a promising concept and its ability to measure spatial self-organization was demonstrated in this thesis. It captures the idea, that for a system to self-organize information needs to spread through the system which results in a correlation of remote parts of the system. As discussed in Section 3.5 and already noticed by Polani (2008), it is 'orthogonal' to the measure of statistical complexity (Crutchfield and Young, 1989) by measuring structure in a spatial dimension (specified by observers) instead of the temporal dimension.

To measure O-organization in the continuous domain, the estimation of multi-information was needed. The estimation of multi-information and mutual information for high

dimensional systems seems to be a rather unexplored field. Though the KSG estimator performed well in these cases (much better than all other estimators), there seems a lack of a good understanding how the bias grows with dimension and no good selection method for k as a parameter.

7.2.2 *Particle & Agent Collectives*

The simulations with particle systems show that for a system with local interactions a certain homogeneity of interactions, here present due to a limited number of different types, leads to a higher level of self-organization. In particular, the relation between organization, interactions being local and their homogeneity is relevant. It seems that hierarchies, as for example induced by the clusters of particles, overcome the limits that local interactions pose on organization. Hierarchies of clusters have another important property and that is a certain robustness to perturbations, which has not been studied here. Ladyman et al. (2013) list robustness and the emergence of hierarchical organization among the properties associated with complex systems. They argue, that robustness is necessary for a complex system. To me it is not entirely clear whether this necessity stems from the structure of most physical models or from a purely information-theoretic or computational mechanics perspective. The results I presented here hint that the structure of interactions plays a more important role for this necessity than the measure of organization.

Particles are passive, the dynamics are determined by the physics alone and there is no local decision making mechanism. If the particles are active agents, I was able to show in Chapter 5 that the amount of self-organization, driven by the agents' actions, is limited by the information processing of the collective. This is relevant, as it is another building block for the general idea of an information book-keeping principle in living organisms (Polani, 2009). Furthermore, it shows that organization is limited by the cognitive capabilities of the collective and the amount the environment and embodiment of the agents works in favour of organization (or disfavour, for example by inducing noise into the system). For example in cell sorting, where the adhesion properties, i.e the embodiment of cells, are enough to drive the whole sorting process (Graner and Glazier, 1992), organization seems to stem only from open-loop control (the cells as embodied agents perform no information processing). This concept is also known under the name of *morphological computation*, a term coined by (Pfeifer and Bongard, 2007). On the other hand, in the process of gastrulation, when the ventral furrow forms (Wolpert et al., 2002), cells need to process positional information (Dubuis et al., 2011) and react by changing local properties. While the exact mechanisms of this are still unclear (Hocevar and Ziherl, 2011), and there are several mechanisms that can drive such a formation, like keystone deformation or change of adhesion properties (Davies, 2005), such mechanism require a minimal amount of cognition in the sense of information processing within the perception-action loop.

I applied the framework of multi-agent relevant information developed in Chapter 5 to a small scenario with two agents. The two agent scenario was designed without the

intent to be biologically plausible or any relation to morphogenesis of living organisms, but serves as a minimal working example of the framework. Furthermore, it shows that intrinsic coordination can overcome limitations for suboptimal agents, where the limitations stem from the mere anticipation of other agents behaviour. To judge, whether intrinsic coordination is more efficient, a further cost needs to be considered. Namely the cost of maintaining an information channel between agents, which in some cases might even be metabolically more expensive than increasing the information processing of each agent. Nonetheless, the framework itself is not limited to such a toy scenario as presented here, even though the underlying algorithms currently limit an easy scaling of the scenario.

7.2.3 *Local & Redundant Information*

In larger collectives it is not only interesting to quantify self-organization, but also to understand the local interaction dynamics between agents, as for example the intrinsic coordination discussed in the last section. In Chapter 4 I used coarse graining of multi-information in an attempt to find type based difference in the self-organization of particle systems. It was possible to distinguish two different phases using this approach, however this assessment did not lead to an insight to local information processing. If the collective consists of active agents instead of passive particles the multi-agent perception-action loop from Chapter 5 can be used to study information processing of the individual agents. For such a purpose the here developed measure of redundancy is fitting.

While the construction of the novel measure of redundant information addressed a problem in the way redundant information was calculated before, there are still some unanswered questions. As of now, there is no measure of redundancy, that fulfils the redundancy axioms by Williams (2011) as well as the left monotonicity axiom (Bertschinger et al., 2012). Comparisons of redundant information values where the target variable Z is of the same dimension are nonetheless possible.

The introduced measure allows to separate information coming from different sources, which together with the decomposition of transfer entropy allows a better understanding of information transfer and storage (Lizier et al., 2013). Furthermore, quantification of redundant information can give information about local 'backup' channels in agent collectives. However redundant information is not just about redundant channels, also mechanisms that do not distinguish from where signals come exhibit redundant information (an example that was shown in Chapter 6 was the AND gate, as well as the operation of addition). On the other hand the quantification of synergy allows to detect mechanisms that 'combine' information, which by some has been considered to be the basis to cognition (Griffith and Koch, 2012).

7.3 FUTURE WORK

7.3.1 *Scaling of Simulations*

If these simulations can be transferred to a more biological correct model, it might be possible to get some results on how physical or chemical laws determine how spatial self-organizing systems necessarily have to be structured and how their dynamics have to look like, purely based on information-theoretic constraints. This requires to have at least a possibility to simulate and analyse larger collectives. Due to computational limitations, the number of particles in the simulations that were carried out could not be increased to a very large number, even using parallelized algorithms on a cluster. It would be very interesting to see how a particle system would scale (leaving the number of types fixed). In such a case, in the equilibrium the size of the whole particle configuration will exceed the cut-off radius by several orders of magnitude.

For the relevant information formalism I carried out some initial tests not reported here, where I successfully implemented relevant information in the continuous domain and believe that using continuous world and kernel based version of the relevant information formalism are the best approach to implement it for larger multi-agent collectives. Though continuous system are usually more complex to deal with computationally, the availability of good estimation methods make them sometimes easier to deploy.

7.3.2 *Extensions of Redundant Information*

If the geometric implications of an extension of the underlying space $\mathcal{Z}$ are better understood, it might be possible to construct a version of the here developed redundancy measure that accounts for the dimension of $\mathcal{Z}$, so that left monotonicity is fulfilled or at least the deviation from it can be accounted for.

Further extensions of the measure include a continuous version, which at the moment poses a similar problem as the multivariate extension with respect to computability. While information projections onto polytopes are easy to compute, this is not the case for arbitrary convex sets.

Moreover, separating the notions of source and mechanistic redundancy, which I hypothesize, need further work. The concepts were only introduced here and are not yet fully explored, I proposed a measure of both quantities and conjectured several relations between the measures. I expect that more work in this direction will lead to a better understanding of the foundations of information theory and its application to information processing in agent collectives.

7.3.3 *Links between Measures of Self-Organization*

I am not aware of a direct link between multi-information of observers of a system and the time dependent version of statistical complexity, however I speculate that it is possible

to relate both quantities. A first hint for this is given by the discussion about the parallel that both measures are maximal in ordered system, where the order is in the temporal or spatial (with respect to observers) dimension. However, the parallel ends here and I am not aware of for example a generalization that could produce both measures as special cases.

7.3.4 *Self-Organization and Biological Transitions*

Shalizi (2001) defined emergence as the existence of a derived process that has higher predictive efficiency and he showed how thermodynamics emerge from statistical mechanics. In a similar way, biological organization, in terms of compartmentalization and hierarchies might emerge from physics as a necessity for self-organizing systems.

The transition from passive particles to active agents seems similar to the transition from physics to biology, the 'origin of life'. How and why this transition happened is still a completely unsettled question and I would guess that there is no purely information-theoretic explanation for it, but rather a physical one. On a more abstract level however, my hypothesis is, that the perception-action loop emerges from the world dynamics (in the sense of emergence by Shalizi (2001)).

There is another important transition in biology, the emergence of signalling pathways, the most prominent example being the nervous system and brains. Hormones and neurotransmitters are other examples. While it is quite hard to argue with evolution at the origin of life, it is much easier at a later stage, when self-replication and evolution is happening in a more well defined way. For example, in the framework of Chapter 5 evolution can account for the coordination of agents in so far, as the genetic code which encodes their policies can anticipate the actions of other agents given their sensor readings. However, in some cases, as the results suggest, it might give a better performance if the actions of the agents of a collective are actually coordinated (not just by merely anticipating the action of another agent). This results in an evolutionary argument why information signalling might have developed.

I believe that these stages, from simple interactions to compartmentalization and hierarchies, then to minimal cognition and in turn coordination and the development of directed ways of information transfer are key to understand the origins of life and the development of individual organisms. Furthermore I am convinced that information-theoretic methods provide a relevant toolbox to study these questions in a quantitative way. The information-theoretic perspective on morphogenesis and agent collectives put forward in this thesis together with the methods developed here, hopefully provide another building block in the general understanding of living organisms as information processing systems.

BIBLIOGRAPHY

Amari, S. and Nagaoka, H. (2007). *Methods of information geometry*, volume 191. Amer Mathematical Society.

Amari, S.-I. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory, 47*(5), 1701–1711.

Anthony, T., Polani, D. and Nehaniv, C. (2008). *On preferred states of agents: how global structure is reflected in local structure*, pages 25-32. MIT Press. Original paper can be found at: http://alifexi.alife.org/proceedings/.

Ashby, W. R. (1956). *An introduction to cybernetics*.. New York, J. Wiley.

Atick, J. J. and Redlich, A. N. (1992). What does the retina know about natural scenes?. *Neural computation, 4*(2), 196–210.

Attneave, F. (1954). Some Informational Aspects of Visual Perception. *Psychological Review, 61*(3), 183–193.

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results*. New York, Holt.

Ay, N., Bertschinger, N., Der, R., Güttler, F. and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B, 63*(3), 329–339.

Ay, N., Olbrich, E., Bertschinger, N. and Jost, J. (2006a). A unifying framework for complexity measures of finite systems. In *Proceedings of ECCS06* .

Ay, N., Olbrich, E., Bertschinger, N. and Jost, J. (2011). A geometric approach to complexity. *Chaos, 21*(3), 037103.

Ay, N. and Polani, D. (2008). Information Flows in Causal Networks. *Advances in Complex Systems, 11*(1), 17–41.

Balduzzi, D., Ortega, P. a. and Besserve, M. (2013). Metabolic Cost As an Organizing Principle for Cooperative Learning. *Advances in Complex Systems*, pp. 1350012.

Balduzzi, D. and Tononi, G. (2008). Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Computional Biology, 4*(6).

Balduzzi, D. and Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS computational biology, 5*(8), e1000462.

Bard, J. (2008). Morphogenesis. *Scholarpedia, 3*(6), 2422.

Bard, J. B. L. (1990). *Morphogenesis : the cellular and molecular processes of developmental anatomy / Jonathan Bard*. Cambridge University Press Cambridge [England] ; New York.

Barlow, H. B. (1959). Possible Principles Underlying the Transformations of Sensory Messages. In Rosenblith, W. A., editor, *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, pages 217–234.

Barlow, H. B. (2001). Redundancy Reduction Revisited. *Network: Computation in Neural Systems, 12*(3), 241–253.

Baylor, D., Lamb, T. and Yau, K.-W. (1979). Responses of retinal rods to single photons.. *The Journal of Physiology, 288*, 613.

Bell, A. (2003). The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003* .

Bellman, R. and Kalaba, R. E. (1965). *Dynamic programming and modern control theory*. Academic Press New York.

Bennett, C. H. (1990). How to define complexity in physics, and why. *Complexity, entropy, and the physics of information, 8*, 137–148.

Bennett, C. H. (1993). Dissipation, information, computational complexity and the definition of organization. *Santa Fe Institute Studies in the Sciences of Complexity - Proceedings Volume 1, 1*, 215.

Bentley, K., Cox, E. J. and Bentley, P. J. (2005). Nature's batik: a computer evolution model of diatom valve morphogenesis. *Journal of nanoscience and nanotechnology, 5*(1), 25–34.

Bertschinger, N., Olbrich, E., Ay, N. and Jost, J. (2008). Autonomy: An information-theoretic perspective. *Biosystems, 91*(2), 331–345.

Bertschinger, N., Rauh, J., Olbrich, E. and Jost, J. (2012). Shared information – new insights and problems in decomposing information in complex systems. *Arxiv preprent, arXiv:1210.5902*.

Bialek, W., Nemenman, I. and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation, 13*(11), 2409–2463.

Biel, M. and Polani, D. (2012). Personal communication. University of Hertfordshire.

Bishop, C. M. and others (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Blahut, R. (1972). Computation of Channel Capacity and Rate Distortion Functions. *IEEE Transactions on Information Theory, 18*(4), 460–473.

Blake, W. J., Kærn, M., Cantor, C. R. and Collins, J. J. (2003). Noise in eukaryotic gene expression. *Nature, 422*(6932), 633–637.

Boltzmann, L. (1866). *Über die mechanische Bedeutung des zweiten Hauptsatzes der Wärmetheorie*. Staatsdruckerei.

Bonabeau, E. (1997). From classical models of morphogenesis to agent-based models of pattern formation. *Artificial Life, 3*(3), 191–211.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press.

Brenner, N., Bialek, W. and de Ruyter van Steveninck, R. (2000). Adaptive rescaling optimizes information transmission. *Neuron, 26*, 695–702.

Capdepuy, P. (2010). *Informational Principles of Perception-Action Loops and Collective Behaviours*. PhD thesis,.

Capdepuy, P., Polani, D. and Nehaniv, C. (2007a). Constructing the basic umwelt of artificial agents. *Lecture Notes in Computer Science*.

Capdepuy, P., Polani, D. and Nehaniv, C. L. (2007b). Maximization of Potential Information Flow as a Universal Utility for Collective Behaviour. *IEEE Symposium on Artificial Life (ALIFE '07)*, pp. 207–213.

Cellucci, C., Albano, A. M. and Rapp, P. (2005). Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical Review E, 71*(6), 066208.

Chaitin, G. J. (1969). On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the ACM (JACM), 16*(3), 407–422.

Chee-Orts, M.-N. and Optican, L. (1993). Cluster method for analysis of transmitted information in multivariate neuronal data. *Biological cybernetics, 69*(1), 29–35.

Cheng, J. (2005). Robust and self-repairing formation control for swarms of mobile agents. In *Proceedings of the Royal Society (London) A 354, 303–330, and 377*, pages 50–4.

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on, 2*(3), 113–124.

Christensen, A. L., O'Grady, R. and Dorigo, M. (2008). SWARMORPH-script: a language for arbitrary morphology generation in self-assembling robots. *Swarm Intelligence, 2*(2), 143–165.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. The MIT Press.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition*. Wiley-Interscience.

Crutchfield, J. (1990). Information and its metric. *Nonlinear Structures in Physical Systems–Pattern Formation, Chaos and Waves*, pp. 119–130.

Crutchfield, J. (1992). Semantics and thermodynamics. In *Santa Fe Institue Studies in the Sciences of Complexity*, pages 317–317.Addison-Wesley Publishing Co
Crutchfield, J. and Young, K. (1989). Inferring statistical complexity. *Physical Review Letters, 63*(July), 105–108.
Crutchfield, J. P. (1994). The calculi of emergence: Computation, dynamics, and induction. *Physica D, 75*, 11–54.
Crutchfield, J. P. and Feldman, D. P. (2001). Regularities unseen, randomness observed: Levels of entropy convergence. *arXiv preprint cond-mat/0102181*.
Crutchfield, J. P. and Shalizi, C. R. (1999). Thermodynamic depth of causal states: Objective complexity via minimal representations. *Physical Review E, 59*(1), 275.
Csiszár, I. and Matus, F. (2003). Information projections revisited. *Information Theory, IEEE Transactions on, 49*(6), 1474–1490.
Csiszar, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial. In Verdú, E.-i.-c. S., Notredame, D. C., Stanford, T. C., Maryland, A. E. and Stanford, A. G., editors, *Foundations and Trends™ in Communications and Information Theory*. .
Dalenoort, G. (1989). *The Paradigm of self-organization*. Gordon and Breach Science Publ..
Darbellay, G. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory, 45*(4), 1315–1321.
Davies, J. (2005). *Mechanisms of morphogenesis*. Academic Press.
Davies, J. (2008). Cellular mechanisms of morphogenesis. *Scholarpedia, 3*(2), 3615.
Davies, R., Twining, C. and Taylor, C. (2008). *Statistical models of shape: optimisation and evaluation*. Springer-Verlag New York Inc.
de Ladurantaye, V., Rouat, J. and Vanden-Abeele, J. (2012). Models of information processing in the visual cortex. .
Denk, W. and Webb, W. W. (1989). Thermal-noise-limited transduction observed in mechanosensory receptors of the inner ear. *Physical Review Letters, 63*(2), 207–210.
Der, R., Steinmetz, U. and Pasemann, F. (1999). Homeokinesis — A new principle to back up evolution with learning. In *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'99)* .
DeWeese, M. and Meister, M. (1999). How to measure the information gained from one symbol. *Network: Computation in Neural Systems, 10*(4), 325–340.
Doursat, R. (2008b). Organically grown architectures: Creating decentralized, autonomous systems by embryomorphic engineering. *Organic Computing*, pp. 167–199.
Doursat, R. (2008a). Programmable Architectures That Are Complex and Self-Organized : From Morphogenesis to Engineering. *Artificial Life*, pp. 181–188.
Dretske, F. (1981). *Knowledge & the flow of information*. Center for the Study of Language and Information, MIT Press.
Dryden, I. and Mardia, K. (1998). *Statistical shape analysis*, volume 4. John Wiley & Sons New York.
Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T. and Bialek, W. (2011). Positional information, in bits. *arXiv preprint arXiv:1201.0198*.
Effenberger, F. (2013). A primer on information theory, with applications to neuroscience. *ArXiv preprint, arXiv:1304.2333*.
Feldman, D. P. and Crutchfield, J. P. (1998). Measures of statistical complexity: Why?. *Physics Letters A, 238*(4), 244–252.
Flecker, B., Alford, W., Beggs, J., Williams, P. and Beer, R. (2011). Partial information decomposition as a spatiotemporal filter. *Chaos, 21*(3).
Floridi, L. (2010). *Information: A very short introduction*. OUP Oxford.
Frank, A. (1980). On chain and antichain families of a partially ordered set. *Journal of Combinatorial Theory, Series B, 29*(2), 176–184.

Fraser, A. and Swinney, H. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A, 33*(2), 1134.
Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. *Probability theory and related fields, 57*(4), 453–476.
Friedman, N., Mosenzon, O., Slonim, N. and Tishby, N. (2006). Multivariate information bottleneck. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 1739–89.
Gallagher, S. (2005). *How the body shapes the mind*. Cambridge Univ Press.
Gat, I. and Tishby, N. (1999). Synergy and redundancy among brain cells of behaving monkeys. *Advances in neural information processing systems*, pp. 111–117.
Gell-Mann, M. and Lloyd, S. (2004). Effective complexity. *Nonextensive entropy*, pp. 387–398.
Gibbs, J. (1874). *On the equilibrium of heterogeneous substances*. Transactions of the Connecticut Academy.
Gibbs, J. W. (2010). *Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics*. Cambridge University Press.
Gibson, J. J. (1986). *The Ecological approach to visual perception*. New edition edition Lawrence Erlbaum Associates.
Gierer, A. and Meinhardt, H. (1972). A theory of biological pattern formation. *Biological Cybernetics, 12*(1), 30–39.
Glackin, C., Salge, C. and Polani, D. (2013). Personal communication. University of Hertfordshire.
Glazier, J. A. and Graner, F. (1993). Simulation of the differential adhesion driven rearrangement of biological cells. *Phys. Rev. E, 47*, 2128–2154.
Graner, F. and Glazier, J. A. (1992). Simulation of biological cell sorting using a two-dimensional extended potts model. *Phys. Rev. Lett., 69*, 2013–2016.
Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics, 25*(9), 907–938.
Gregor, T., Tank, D. W., Wieschaus, E. F. and Bialek, W. (2007). Probing the limits to positional information. *Cell, 130*(1), 153.
Griffith, V. (2011). Quantifying synergistic information remains an unsolved problem. *Arxiv preprint arXiv:1122.1680v3*.
Griffith, V. and Koch, C. (2012). Quantifying synergistic mutual information. *Arxiv preprint arXiv:1205.4265*.
Grinspun, E., Desbrun, M., Polthier, K., Schröder, P. and Stern, A. (2006). Discrete differential geometry: an applied introduction. *ACM SIGGRAPH Course*.
Gumbiner, B. M. and others (1996). Cell adhesion: the molecular basis of tissue architecture and morphogenesis.. *Cell, 84*(3), 345–357.
Hansen, E. A. (2008). Sparse stochastic finite-state controllers for pomdps.. In *UAI*, pages 256–263.
Harder, M. and Polani, D. (2012). Self-organizing particle systems. *Advances in Complex Systems*, pp. 1250089.
Harder, M., Polani, D. and Nehaniv, C. (2011). Think globally, sense locally: From local information to global features. In *Artificial Life (ALIFE), 2011 IEEE Symposium on*, pages 70–77.
Harder, M., Polani, D. and Nehaniv, C. L. (2010). Two agents acting as one. In Fellermann, H., Dörr, M., Hanczyc, M., Ladegaard, L. L. and Maurer, S. et al., editors, *Artificial Life XII: The 12th International Conference on the Synthesis and Simulation of Living Systems*, pages 599-606.
Harder, M., Salge, C. and Polani, D. (2013). Bivariate measure of redundant information. *Physical Review E, 87*(1), 012130.
Harrison, L. G. (1994). Kinetic theory of living pattern.. *Endeavour, 18*(4), 130–136.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research, 10*, 1469–1484.

Hecht, S., Shlaer, S. and Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of General Physiology, 25*(6), 819–840.

Herzel, H. and Groe, I. (1995). Measuring correlations in symbol sequences. *Physica A: Statistical Mechanics and its Applications, 216*(4), 518–542.

Hocevar, A. and Ziherl, P. (2011). Collective mechanics of embryogenesis: Formation of ventral furrow in Drosophila. *arXiv preprint arXiv:1108.4795*, pp. 3–6.

Hogeweg, P. (2000). Shapes in the shadow: evolutionary dynamics of morphogenesis. *Artificial Life, 6*(1), 85–101.

Houchmandzadeh, B., Wieschaus, E. and Leibler, S. (2002). Establishment of developmental precision and proportions in the early drosophila embryo. *Nature, 415*(6873), 798–802.

Hulle, M. M. V. (2005). Edgeworth approximation of multivariate differential entropy. *Neural Computation, 17*(9), 1903–1910.

Jade, S. and Sarkar, S. (1993). Statistical models for slope instability classification. *Engineering Geology, 36*(1), 91–98.

Jakobsson, L., Franco, C. A., Bentley, K., Collins, R. T. and Ponsioen, B. et al. (2010). Endothelial cells dynamically compete for the tip cell position during angiogenic sprouting. *Nature cell biology, 12*(10), 943–953.

James, R., Ellison, C. and Crutchfield, J. (2011). Anatomy of a bit: Information in a time series observation. *Arxiv preprint arXiv:1105.2988*.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M. and Schoelkopf, B. (2012). Quantifying causal influences. *ArXiv preprint, arXiv:1203.6502*.

Jaynes, E. (1992). *The Gibbs paradox*. Kluwer Academic.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106(4)*, 620–630.

Jones, J. (2010). Characteristics of pattern formation and evolution in approximations of physarum transport networks.. *Artificial Life, 16*(2), 127–53.

Kahle, T., Olbrich, E., Jost, J. and Ay, N. (2009). Complexity measures from interaction structures. *Physical Review E, 79*(2), 026201.

Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S. and Erickson, D. J. et al. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data.. *Physical Review E, 76*(2), 026209.

Klenke, A. (2008). *Probability theory: a comprehensive course*. Springer Verlag London.

Kloeden, P., Platen, E. and Schurz, H. (1994). Stochastic differential equations. *Numerical Solution of SDE Through Computer Experiments*, pp. 63–90.

Klyubin, A. S., Polani, D. and Nehaniv, C. L. (2004). Tracking Information Flow through the Environment: Simple Cases of Stigmergy. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T. and Watson, R. A., editors, *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pages 563–568. Cambridge, MA.

Klyubin, A. S., Polani, D. and Nehaniv, C. L. (2007). Representations of Space and Time in the Maximization of Information Flow in the Perception-Action Loop. *Neural Computation, 19*(9), 2387–2432.

Knabe, J., Schilstra, M. and Nehaniv, C. (2008). Evolution and morphogenesis of differentiated multicellular organisms: autonomously generated diffusion gradients for positional information. , pp. 321–328.

Kolasa, J. (2005). Complexity, system integration, and susceptibility to change: Biodiversity connection. *Ecological Complexity, 2*(4), 431 - 442.

Kolmogorov, A. N. (1946). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Chelsea Pub. Co.
Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A, 25*(4), 369–376.
Kondor, I. (2008). *Group theoretical methods in machine learning*. Columbia University.
Kozachenko, L. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii, 23*(2), 9–16.
Kraskov, A., Stögbauer, H. and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E, 69*, 066138.
Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium*, pp. 481–492.
Ladyman, J., Lambert, J. and Wiesner, K. (2011). What is a complex system?. .
Ladyman, J., Lambert, J. and Wiesner, K. (2013). What is a complex system?. *European Journal for Philosophy of Science, 3*(1), 33–67.
Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM journal of research and development, 5*(3), 183–191.
Landauer, R. (1991). Information is physical. *Physics Today, 44*, 23.
Latham, P. E. and Nirenberg, S. (2005). Synergy, redundancy, and independence in population codes, revisited. *The Journal of neuroscience, 25*(21), 5195–5206.
Laughlin, S. B., de Ruyter van Steveninck, R. R. and Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience, 1*(1), 36–41.
Lee, J., Nemati, S., Silva, I., Edwards, B. a. and Butler, J. P. et al. (2012). Transfer entropy estimation and directional coupling change detection in biomedical time series. *Biomedical engineering online, 11*(1), 19.
Lee, J. M. (1997). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Verlag.
Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434.
Lehn, J.-M. (2002). Toward self-organization and complex matter. *Science, 295*(5564), 2400–2403.
Li, M. and Vitanyi, P. M. (1993). *An introduction to Kolmogorov complexity and its applications*. Springer Verlag.
Liang, K. and Wang, X. (2008). Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology, 2008*.
Liggett, T. (1985). *Interacting particle systems*, volume 276. Springer Verlag.
Linsker, R. (1988). Self-organization in a perceptual network. *Computer, 21*(3), 105–117.
Lizier, J. (2011). Information Storage and Transfer in the Synchronization Process in Locally-Connected Networks. *Phd Thesis*.
Lizier, J., Prokopenko, M. and Zomaya, A. (2008). Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E, 77*(2), 026110.
Lizier, J. T., Flecker, B. and Williams, P. L. (2013). Towards a synergy-based approach to measuring information modification. *arXiv preprint arXiv:1303.3440*.
Lizier, J. T., Prokopenko, M. and Zomaya, A. Y. (2007). Information transfer by particles in cellular automata. In *Progress in Artificial Life*, pages 49–60. Springer.
Lloyd, S. and Pagels, H. (1988). Complexity as thermodynamic depth. *Annals of Physics, 188*(1), 186–213.
Lwoff, A. (1962). *Biological order*, volume 26. MIT press Cambridge, Massachusetts.
MacMahon, J. A., Phillips, D. L., Robinson, J. V. and Schimpf, D. J. (1978). Levels of biological organization: an organism-centered approach. *BioScience*, pp. 700–704.
Marée, A. F. and Hogeweg, P. (2001). How amoeboids self-organize into a fruiting body: multicellular coordination in Dictyostelium discoideum.. *Proceedings of the National Academy of Sciences of the United States of America, 98*(7), 3879–83.

Margolin, A. a., Nemenman, I., Basso, K., Wiggins, C. and Stolovitzky, G. et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics, 7 Suppl 1*, 7.

Maurer, U. and Wolf, S. (1999). Unconditionally secure key agreement and the intrinsic conditional information. *Information Theory, IEEE Transactions on, 45*(2), 499 -514.

McGill, W. (1954). Multivariate information transmission. *Information Theory, IRE Professional Group on, 4*(4), 93–111.

Meinhardt, H. (1982). *Models of biological pattern formation*. Academic Press London.

Meinhardt, H. (2006). Primary body axes of vertebrates: generation of a near-Cartesian coordinate system and the role of Spemann-type organizer.. *Developmental Dynamics, 235*(11), 2907–2919.

Miller, G. A. (1955). Note on the bias of information estimates. *Information theory in psychology: Problems and methods, 2*, 95–100.

Miller, J. (2004). Evolving a self-repairing, self-regulating, French flag organism. *Genetic and Evolutionary Computation – GECCO*.

Mondada, F., Pettinaro, G. C., Guignard, A., Kwee, I. W. and Floreano, D. et al. (2004). Swarm-bot: A new distributed robotic concept. *Autonomous Robots, 17*(2-3), 193–221.

Moon, Y., Rajagopalan, B. and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Physical Review E, 52*(3), 2318.

Nemenman, I., Shafee, F. and Bialek, W. (2002). Entropy and inference, revisited. *Arxiv preprint, arXiv:0108:025v2*.

Odell, G., Oster, G., Burnside, B. and Alberch, P. (1980). A mechanical model for epithelial morphogenesis. *Journal of Mathematical Biology, 9*(3), 291–295.

Olsson, L., Nehaniv, C. and Polani, D. (2004). Sensory channel grouping and structure from uninterpreted sensor data. In *Evolvable Hardware, 2004. Proceedings. 2004 NASA/DoD Conference on*, pages 153–160.IEEE

Olsson, L., Nehaniv, C. L. and Polani, D. (2005). Sensor Adaptation and Development in Robots by Entropy Maximization of Sensory Data. In *Proceedings of the 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA 2005). IEEE Computer*, pages 587–592.

Optican, L., Gawne, T., Richmond, B. and Joseph, P. (1991). Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biological cybernetics, 65*(5), 305–310.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature genetics, 31*(1), 69–73.

Panzeri, S. and Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems, 7*(1), 87–107.

Papana, A. and Kugiumtzis, D. (2008). Evaluation of mutual information estimators on nonlinear dynamic systems. *arXiv preprint arXiv:0809.2149*.

Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge University Press.

Pfeifer, R. and Bongard, J. (2007). *How the body shapes the way we think: a new view of intelligence*. Bradford Books.

Pfeifer, R. and Scheier, C. (2001). *Understanding intelligence*. The MIT Press.

Polani, D. (2002). Measures for the organization of self-organizing maps. *Studies in Fuzziness and Soft Computing, 78*, 13–44.

Polani, D. (2004). Defining emergent descriptions by information preservation. *Proc. of the International Conference on Complex*.

Polani, D. (2006). Emergence, Intrinsic Structure of Information, and Agenthood. In .

Polani, D. (2008). Foundations and formalizations of self-organization. *Advances in applied self-organizing systems*, pp. 19–37.

Polani, D. (2009). Information: currency of life?. *HFSP Journal, 3*(5), 307–316.
Polani, D. (2011). An informational perspective on how the embodiment can relieve cognitive burden. In *Artificial Life (ALIFE), 2011 IEEE Symposium on*, pages 78–85.IEEE
Polani, D., Nehaniv, C. L., Martinetz, T. and Kim, J. T. (2006). Relevant Information in Optimized Persistence vs. Progeny Strategies. In *Artificial Life X : Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, pages 337–343.
Press, W., Flannery, B., Teukolsky, S., Vetterling, W. and others (1986). *Numerical recipes*, volume 547. Cambridge University Press.
Prusinkiewicz, P. (1993). Visual models of morphogenesis. *Artif. Life, 1*(1-2), 61–74.
Quastler, H. (1964). *The emergence of biological organization*. Yale University Press New Haven.
Rashid, A. M., Albert, I., Cosley, D., Lam, S. K. and McNee, S. M. et al. (2002). Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134.ACM
Rosa, M. D., Goldstein, S. C., Lee, P., Campbell, J. D. and Pillai, P. (2008). Programming modular robots with locally distributed predicates. In *Proceedings of the IEEE ICRA*, pages 3156–3162.
Rubenstein, M., Ahler, C. and Nagpal, R. (2012). Kilobot: A low cost scalable robot system for collective behaviors. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3293–3298.
Rusu, R. and Cousins, S. (2011). 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4.
Saerens, M., Achbany, Y., Fouss, F. and Yen, L. (2009). Randomized shortest-path problems: Two related models. *Neural Computation, 21*(8), 2363–2404.
Salge, C., Glackin, C. and Polani, D. (2013). Approximation of empowerment in the continuous domain. *Advances in Complex Systems, 16*(1 & 2).
Salge, C. and Polani, D. (2009). Information-theoretic Incentives for Social Interaction. *Technical Report 495, University of Hertfordshire, Hatfield*.
Sayama, H. (2009). Swarm chemistry. *Artificial life, 15*(1), 105–14.
Schinazi, R. B. (1999). *Classical and spatial stochastic processes*. Birkhäuser Boston.
Schreiber, T. (2000). Measuring information transfer. *Physical review letters, 85*(2), 461–464.
Scott, D. W. and Sain, S. R. (2005). Multi-dimensional density estimation. *Handbook of Statistics, 24*, 229–261.
Shalizi, C. R. (2001). *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. PhD thesis,.
Shalizi, C. R. and Crutchfield, J. P. (2002). Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction. *Advances in Complex Systems, 5*(01), 91–95.
Shalizi, C. R. and Klinkner, K. L. (2004). Blind construction of optimal nonlinear recursive predictors for discrete sequences. In Chickering, M. and Halpern, J. Y., editors, *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)*, pages 504–511. Arlington, Virginia.
Shalizi, C. R. and Shalizi, K. (2003). Optimal nonlinear prediction of random fields on networks.. In *DMCS*, pages 11–30.
Shalizi, C. R., Shalizi, K. L. and Haslinger, R. (2004). Quantifying self-organization with optimal predictors. *Physical Review Letters, 93*(11), 118701.
Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(October), 379–423.

Shepard, R. N. (1984). Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review, 91*(4), 417–447.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC.
Sinnott, E. (1960). *Plant Morphogenesis*. Mcgraw-hill Book Company.
Slonim, N., Atwal, G. S., Tkacik, G. and Bialek, W. (2005). Estimating mutual information and multi-information in large networks. *CoRR, abs/cs/0502017*.
Small, C. (1996). *The statistical theory of shape*. Springer Verlag.
Steudel, B. and Ay, N. (2010). Information-theoretic inference of common ancestors. *Arxiv preprint arXiv:1010.5720*.
Steuer, R., Kurths, J., Daub, C. O., Weise, J. and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables.. *Bioinformatics (Oxford, England), 18 Suppl 2*, 231.
Still, S. (2009). Information-theoretic approach to interactive learning. *EPL (Europhysics Letters), 85*, 28005.
Still, S., Crutchfield, J. P. and Ellison, C. J. (2007). Optimal causal inference. *CoRR, abs/0708.1580*.
Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. and Bialek, W. (1996). Entropy and Information in Neural Spike Trains. *eprint arXiv:cond-mat/9603127*.
Summerbell, D., Lewis, J. and Wolpert, L. (1973). Positional information in chick limb morphogenesis. *Nature, 244*, 492–496.
Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press.
Suzuki, T., Sugiyama, M., Sese, J. and Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. In *JMLR workshop and conference proceedings*, pages 5–20.
Theraulaz, G. and Bonabeau, E. (1995). Modelling the collective building of complex architectures in social insects with lattice swarms. *Journal of Theoretical Biology, 177*(4), 381–400.
Thom, R. (1989). *Structural stability and morphogenesis*. Addison Wesley Publishing Company.
Thompson, D. and Bonner, J. (1992). *On Growth and Form*. Cambridge University Press.
Tickle, C., Summerbell, D. and Wolpert, L. (1975). Positional signalling and specification of digits in chick limb morphogenesis. *Nature, 254*(5497), 199–202.
Tishby, N., Pereira, F. C. and Bialek, W. (1999). The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.
Tishby, N. and Polani, D. (2010). Information Theory of Decisions and Actions. In Cutsuridis, V., Hussain, A. and Taylor, J., editors, *Perception-Reason-Action Cycle: Models, Algorithms and Systems.* .
Tkačik, G., Callan Jr, C. G. and Bialek, W. (2008). Information capacity of genetic regulatory elements. *Physical Review E, 78*(1), 011910.
Tkačik, G., Walczak, A. M. and Bialek, W. (2009). Optimizing information flow in small genetic networks. *Physical Review E, 80*(3), 031920.
Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences, 106*(28), 11478–11483.
Tononi, G., Sporns, O. and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences, 91*(11), 5033–5037.
Topp, C. F., Wang, W., Cloy, J. M., Rees, R. M. and Hughes, G. (2013). Information properties of boundary line models for n2o emissions from agricultural soils. *Entropy, 15*(3), 972–987.

Touchette, H. and Lloyd, S. (2000). Information-theoretic limits of control. *Physical review letters, 84*(6), 1156–1159.
Touchette, H. and Lloyd, S. (2004). Information-theoretic approach to the study of control systems. *Physica A: Statistical Mechanics and its Applications, 331*(1), 140–172.
Townes, P. L. and Holtfreter, J. (1955). Directed movements and selective adhesion of embryonic amphibian cells. *Journal of experimental zoology, 128*(1), 53–120.
Treves, A. and Panzeri, S. (1995). The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Computation, 7*(2), 399–407.
Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society, 42*(2), 230–265.
Turing, A. M. (1952). The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 237*(641), 37–72.
van Dijk, S. G. and Polani, D. (2012). Informational drives for sensor evolution. In *Artificial Life*, pages 333–340.
van Dijk, S. G., Polani, D. and Nehaniv, C. L. (2010). What do You Want to do Today? Relevant-Information Bookkeeping in Goal-Oriented Behaviour. In Fellermann, H., Dörr, M., Hanczyc, M., Ladegaard, L. L. and Maurer, S. et al., editors, *Artificial Life XII: The 12th International Conference on the Synthesis and Simulation of Living Systems*, pages 176–183. Odense, Denmark.
Vasiev, B., Balter, A., Chaplain, M., Glazier, J. A. and Weijer, C. J. (2010). Modeling gastrulation in the chick embryo: formation of the primitive streak. *PLoS One, 5*(5), e10571.
Vergassola, M., Villermaux, E. and Shraiman, B. I. (2007). 'Infotaxis' as a strategy for searching without gradients. *Nature*, pp. 406–409.
Victor, J. D. (2000). Asymptotic bias in information estimates and the exponential (Bell) polynomials.. *Neural computation, 12*(12), 2797–804.
Von Dassow, G., Meir, E., Munro, E. M. and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature, 406*(6792), 188–192.
Vu, V., Yu, B. and Kass, R. (2007). Coverage-adjusted entropy estimation. *Statistics in medicine, 26*(21), 4039–4060.
Wang, X. R., Miller, J. M., Lizier, J. T., Prokopenko, M. and Rossi, L. F. (2011). Measuring information storage and transfer in swarms. In *Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems*, pages 838–845.
Weng, G., Bhalla, U. S. and Iyengar, R. (1999). Complexity in biological signaling systems. *Science, 284*, 92–96.
Werfel, J. and Nagpal, R. (2006). Extended Stigmergy in Collective Construction. *IEEE Intelligent Systems, 21*(2), 20–28.
Wiener, N. (1948). *Cybernetics; or control and communication in the animal and the machine*. John Wiley.
Williams, P. L. (2011). *Information Dynamics: Its Theory and Application to Embodied Cognitive Systems*. PhD thesis, Indiana University.
Williams, P. L. and Beer, R. D. (2010). Nonnegative Decomposition of Multivariate Information. *Arxiv preprint arXiv:1004.2515*.
Williams, P. L. and Beer, R. D. (2011). Generalized Measures of Information Transfer. *Arxiv preprint arXiv:1102.1507*.
Witten Jr, T. and Sander, L. (1981). Diffusion-limited aggregation, a kinetic critical phenomenon. *Physical Review Letters, 47*(19), 1400–1403.
Wolfram, S. (1986). Theory and applications of cellular automata. *Advanced Series on Complex Systems*.
Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology, 25*(1), 46.

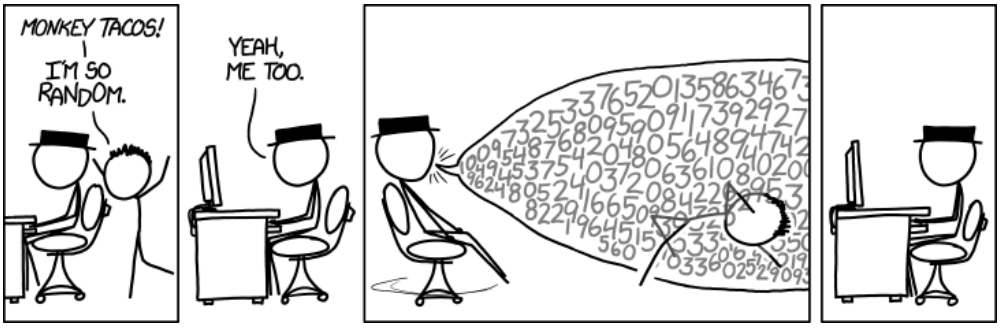
Wolpert, L., Beddington, R., Jessell, T., Lawrence, P. and Meyerowitz, E. et al. (2002). *Principles of development*, volume 3. Oxford University Press New York.:.

Yeung, R. W. (2008). *Information theory and network coding*. Springer.

Yu, C.-H. and Nagpal, R. (2011). A self-adaptive framework for modular robots in a dynamic environment: theory and applications. *The International Journal of Robotics Research, 30*(8), 1015–1036.

Zahedi, K., Ay, N. and Der, R. (2009). Higher coordination with less control - A result of information maximisation in the sensori-motor loop. *CoRR, abs/0910.2*.

Zhang, Z. (1992). Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision, 13*(2), 119–152.



» *In retrospect, it's weird that as a kid I thought completely random outbursts made me seem interesting, given that from an information theory point of view, lexical white noise is just about the opposite of interesting by definition.* «

XKCD, 1210