# INFORMATION DYNAMICS: ITS THEORY AND APPLICATION TO EMBODIED COGNITIVE SYSTEMS

Paul L. Williams

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Cognitive Science Program,
Indiana University

September 2011

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Randall D. Beer, Ph.D.
(Chair)

John M. Beggs, Ph.D.

Olaf Sporns, Ph.D.

September 14, 2011

Larry S. Yaeger

Paul L. Williams

Information Dynamics: Its Theory and Application to Embodied Cognitive Systems

In this thesis, I develop a novel framework for quantifying the information dynamics of brain-body-environment systems, and demonstrate its application to several model agents. I begin by extending one of the core concepts of information theory—the mutual information—into a new method for decomposing informational relationships between multiple variables, called partial information decomposition. Then, using partial information decomposition as the theoretical backbone, I derive techniques for quantifying the flow of information both within and between the components of a brain-body-environment system. Finally, I apply these techniques to analyze specific examples of embodied cognitive behavior, analyzing model agents that perform a simple relational comparison task between two visually presented objects. I explore questions such as how these agents extract and store information about individual objects, and how they integrate information about different objects. Among other key findings, this analysis shows how agents utilize their bodies to actively elicit and structure information from their environments, and use their bodies as extra degrees of freedom to store and process information, illustrating some of the unique ways that embodiment can influence intelligent behavior.

# Contents

# List of Figures

# 1

# Introduction

The aim of this thesis is to develop an information-theoretic framework for the analysis of embodied cognitive systems. By embodied cognitive system, we mean a neural network coupled to a body, which in turn is coupled to an environment, such that together this brain-body-environment system produces adaptive behavior. The dynamics of such a system refers to the time-dependent activity of its individual components—neural, bodily, and environmental state variables—and the network of interactions between its components that collectively give rise to its behavior. Our goal is to understand this dynamics in terms of how information about particular stimulus features flows through the brain-body-environment system. To achieve this goal, three key challenges must be met:

1. Mutual information, one of the core concepts of information theory, must be generalized to deal properly with multivariate interactions. This is necessary because the quantities of interest for information dynamics are fundamentally multivariate in nature. For example, in the simplest case of information flow for a single component, one must consider at least three variables: one representing the stimulus and two representing the states of the component at different times. Meanwhile, there are several known issues with existing multivariate forms of mutual information.

2. With multivariate information as the theoretical foundation, techniques must be developed to quantify various aspects of information dynamics, including how information flows within components, how information is gained and lost by components, and how information is transferred between components.

3. To evaluate these techniques, they must be applied to specific examples of embodied cognitive systems. These examples must be complex enough to raise interesting analytical questions while remaining simple enough to be analytically tractable.

The next section of this chapter provides an overview of the thesis, explaining how it will address each of the three key challenges identified above. The following section then provides a detailed description of how the thesis is organized.

## 1.1 Overview

Following introductory matters, the first part of this thesis (Chapter 4) develops a novel solution to the problem of quantifying multivariate information. Our approach begins with an axiomatic definition of redundancy as the overlapping information that several information sources share about a given variable. Then, we show how this definition of redundancy induces a lattice over sets of information sources that clarifies the general structure of multivariate information. Applying a form of inclusion-exclusion over this lattice, we then derive a definition of partial information (PI) atoms that exhaustively decompose the information in a multivariate system. We show that each of these PI-atoms supports a clear interpretation as an informational quantity, representing the redundancy between synergies for a collection of sources. Finally, comparing PI-decomposition with interaction information—the current de facto measure of multivariate information—we show that several confusing properties of interaction information can be explained by its confounding of redundancy and synergy.

In the second part of this thesis (Chapter 5), we apply PI-decomposition to develop a toolkit for information dynamics. This application requires a shift from thinking about random variables to stochastic processes, or time-indexed sequences of random variables. Specifically, we adopt the view of a brain-body-environment system as a set of coupled stochastic processes, one for each component of the system, with the aim of understanding how information flows through these processes.

We begin by using PI-decomposition to extend transfer entropy—the standard measure of information transfer between stochastic processes—in two complementary ways. First, we decompose transfer entropy into distinct measures of state-dependent and state-independent transfer, based on whether the influence of a source process depends on the state of the target process. As we show, the distinction between these two measures is formally analogous to the fundamental distinction in control theory between closed-loop (state-dependent) and open-loop (state-independent) control. Second, we apply a similar decomposition to the case of multiple interacting processes to derive a novel multivariate generalization of transfer entropy. The resulting measures quantify separately the unique, redundant, and synergistic influences of multiple sources onto a target.

Next, we introduce techniques for quantifying the flow of information *about* some stimulus, which forms a subset of the complete information flow. The flow of stimulus-specific information for an individual component is defined as the mutual information between that component and the stimulus as a function of time. To characterize this flow, we define measures of information gain, loss, and transfer. The first two of these measures capture how the information carried by a component changes from one moment to the next. Specifically, the information carried by a component at time $t$ equals the information that it carried at time $t-1$, minus the information loss and plus the information gain. The third measure, information transfer, is similar in spirit to transfer entropy but quantifies only the information that is transferred *about* some stimulus. This definition of information transfer builds on the definition of information gain, and can be thought of as answering the question: Of the new information gained by a target process, how much is shared (provided redundantly) by a source at the previous time step?

Finally, in the third part of this thesis (Chapters 6 and 7), we apply our information dynamics toolkit to analyze specific examples of embodied cognitive systems. The systems that we consider are simulation models of minimally cognitive behavior, chosen to balance the opposing requirements of behavioral richness and analytical tractability. Specifically, we analyze model agents that were evolved using genetic algorithms to perform a simple form of relational categorization. The task faced by these agents is to categorize objects based on the relational property *smaller than*. In each trial of the task, an agent is presented with two differently sized objects that fall towards it one after the other, with the objective of catching the second object if it is smaller than the first and avoiding it otherwise.

Using the tools of information dynamics, we first examine how these agents extract and store information about the size of the first object, so that it can later be compared with that of the second object. In particular, we begin by quantifying the information stored by all neural and bodily state variables at the instant when the first object is removed. Then, using the techniques described above, we work backwards to determine the origins of this information in the sensory stream. Second, we analyze how agents integrate information about the sizes of the two objects in order to make their relational discrimination. This analysis focuses on the time period during which the second object is presented, and proceeds by first quantifying the flow of information for each object individually, then relating these to the flow of information about the relative size of the two objects. We then conclude by exploring some of the specific ways that embodiment influences the dynamics of information flow in certain of the relational agents. We show evidence that agents both offload information to their environments and use their bodies to actively structure the information that they receive, and furthermore that these informational consequences of embodiment can be characterized in a rigorous and quantitative manner using the tools of information dynamics.

## 1.2 Thesis Organization

The rest of this dissertation is organized as follows. Chapters 2 and 3 lay the foundation for our research contributions by presenting the requisite introductory and background material. Chapter 2 reviews basic concepts from information theory, beginning with the fundamental measures introduced by Shannon—the entropy and mutual information—and proceeding to several other information-theoretic measures developed more recently, including dynamic measures such as the transfer entropy. Then, in Chapter 3, we provide an overview of information-theoretic applications in neuroscience, complex systems, and embodied cognitive science that relate to our framework. As described later, our framework is distinguished by its use of *multivariate*, *dynamic*, and *specific* informational measures that apply to *embodied* systems. Consequently, the overview in Chapter 3 focuses on work that incorporates one or more of these key features of our framework.

Chapter 4 marks the beginning of the research contributions of this thesis. Here we develop the theoretical foundation for our approach to information dynamics, a new method for decomposing the information in a multivariate system called PI-decomposition. The chapter begins with a discussion of previous attempts to extend mutual information to multivariate interactions. Then, we develop the method of PI-decomposition, beginning from a first principles analysis of the structure of multivariate information. Finally, we discuss how PI-decomposition relates to interaction information, and show that our method resolves several longstanding confusions with this influential measure of multivariate information. This chapter is based on work in collaboration with Randall Beer [250].

In Chapter 5, we apply PI-decomposition to develop two sets of techniques for quantifying information dynamics. First, we develop techniques for quantifying the flow of *intrinsic* information, or the flow of information within a stochastic process $X$ where the informational quantity of interest is a property of $X$ itself. The canonical measure of this kind is the transfer entropy, which quantifies the transfer of information into a process $X$ that is about the future state of $X$. Indeed, the techniques developed in the first part of this chapter are best thought of as extensions to transfer entropy, allowing one to tease apart intrinsic information flow in ways that transfer entropy does not afford. Second, we develop techniques for quantifying the flow of *extrinsic* information, or the flow of information within and between stochastic processes that is about something external to those processes. These techniques are later put into action in Chapter 7, when we use them to explore how information about external stimulus features flows through the components of brain-body-environment systems. This chapter is based on work in collaboration with Randall Beer [251].

Chapter 6 introduces the model agents that we will use to illustrate the application of information dynamics to embodied cognitive systems. These agents perform a simple kind of relational categorization, which is a phenomenon that has garnered considerable interest in cognitive science and poses a number of challenging problems for analysis. The chapter begins with a brief overview of relational categorization and its importance for cognitive science. Then we describe previous efforts to model relational categorization and elaborate on some of the important features that distinguish our modeling approach. Afterwards, we detail the methods used to evolve our model agents and present the results from a set of evolutionary experiments. Finally, we present an analysis of the best evolved agent using the mathematical tools of dynamical systems theory, with this analysis providing some preliminary insights into the mechanisms underlying relational categorization in our model agents. This chapter is based on work in collaboration with Randall Beer and

Michael Gasser [252].

In Chapter 7, we apply our information dynamics framework to analyze several features of the relational agents. The chapter begins with a detailed explanation of our analytical approach, describing how the tools introduced in Chapter 5 are applied to our model agents. Then, we analyze several features of the agent that was the focus of Chapter 6, including how the agent extracts information about the sizes of objects and how the agent integrates information about the sizes of different objects. Afterwards, we analyze a second agent that uses a more active behavioral strategy to perform the categorization task. The analysis of this agent illustrates several of the unique ways that embodiment can influence information dynamics, including the ability of embodied agents to actively elicit and structure sensory information, and the ability to offload information to their bodies and environments. This chapter is based on work in collaboration with Randall Beer [249].

Finally, in Chapter 8 we conclude the thesis by summarizing its main contributions and identifying several promising avenues for future research.

# 2

# Information-Theoretic Foundations

In a landmark paper of 1948, Claude Shannon launched the field of information theory while simultaneously formulating and answering its two most fundamental questions ( [206], see also [36,146]). The two questions posed by Shannon were:

(i) What is the shortest possible encoding for a data source such that it can be recovered without error?

(ii) What is the maximum rate at which data can be transmitted error-free over a noisy channel?

The answer to (i), proven in Shannon's source coding theorem, is a fundamental quantity called the *entropy*. The answer to (ii), proven in Shannon's channel coding theorem, is a measure called the channel capacity, which is defined in terms of a second fundamental quantity called the *mutual information*. With these two results, Shannon thus established the absolute limits for lossless compression and lossless transmission, respectively.

Shortly after the publication of these results, the significance of Shannon's ideas was recognized and enthusiastically embraced by researchers in a diverse array of fields. Indeed, the enthusiasm was so great that Shannon himself felt compelled to caution against the indiscriminate application of information theory to areas beyond its original purview [205]. This enthusiasm was rooted in the fact that the fundamental quantities identified by Shannon—the entropy and mutual information—provide extremely general measures of the uncertainty associated with individual variables and of the interdependence between variables, respectively. In contrast with related concepts from classical statistics, these measures have the advantages that they are sensitive to arbitrary relationships between variables—linear or nonlinear, stochastic or deterministic—and have units of measurement (bits) that are easily interpreted and compared across systems. Consequently, following the initial wave of enthusiasm, information theory has continued to grow in significance as a general tool for the analysis of complex systems, with applications in cognitive science, neuroscience, genetics, physics, machine learning, and many other areas.

In this chapter, we begin by discussing the fundamental measures introduced by Shannon, and then proceed to describe several other information-theoretic measures developed more recently. Thus, this chapter will lay the mathematical groundwork both for our discussion of existing applications of information theory (Chapter 3) and for the novel contributions of this thesis (which begin in Chapter 4).

## 2.1 Entropy

The *entropy* for a discrete[1] random variable $X$ with probability mass function $p(x)$ is defined by

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)}, \tag{2.1}$$

where the logarithm is conventionally taken to the base $2$, yielding units of bits. As mentioned above, the concept of entropy originated with Shannon's source coding theorem [206], which established the absolute lower bound on lossless data compression.

[1] Entropy can also be extended to continuous variables, and likewise for the other measures introduced in this chapter. Essentially, one just replaces sums with integrals. However, we will only need the discrete forms throughout this thesis, since whenever continuous variables arise they will be quantized.

This theorem considers the problem of assigning codewords to sequences of $n$ independent samples of $X$—thought of as messages originating from a source $X$—such that each sequence is assigned a unique codeword and the average code length is minimized. This minimization is achieved by assigning codewords such that the length of each is inversely related to the probability of its message, so that on average shorter codewords occur more frequently than longer ones. Shannon's source coding theorem states that, in the $n \to \infty$ limit, the shortest possible encoding requires an average of $H(X)$ bits per outcome of $X$. Consequently, $H(X)$ has the operational meaning that it is the average length of the shortest possible description of $X$, and for this reason $H(X)$ is often interpreted as a measure of the *information content* of $X$[2].

Complementary to its interpretation as information content, entropy is also commonly viewed as the amount of *uncertainty* associated with a random variable. More precisely, entropy corresponds to a measure of statistical dispersion for nominal distributions. Thus, the entropy $H(X)$ is maximal when all outcomes of $X$ are equally probable—representing complete uncertainty—and takes its minimum value of zero when a single outcome occurs with certainty. For example, the entropy for a binary variable representing a coin flip attains its maximum value of one bit if the coin is fair, and decreases to zero as the coin becomes increasingly biased towards heads or tails. In fact, Shannon [206] proved that entropy is the *only* measure of uncertainty that satisfies certain simple and plausible axioms. Thus, in this sense, entropy can be considered the definitive measure of uncertainty, being uniquely specified out of the space of all possible measures. The axioms that Shannon postulated were:

(i) Continuity: A small change in the probability of any outcome should result in a small

[2] This notion of information content as the shortest possible description for an object was later extended in algorithmic information theory [6, 127], where the *algorithmic entropy* of an object, also called its Kolmogorov complexity, is the length of its shortest description in some universal description language (e.g., Turing machine programs). For random variables, there is a deep connection between the Shannon and algorithmic entropies [36].

change in uncertainty.

(ii) Monotonicity: If all outcomes are equiprobable, uncertainty should increase monotonically with the number of outcomes. With equiprobable outcomes, there is more uncertainty with more possible outcomes.

(iii) Additivity: The amount of uncertainty should be independent of any way of grouping outcomes into a succession of choices.

Subsequently, a number of alternative axiomatizations of entropy have also been proposed [42,43,117,150,157].

A third interpretation of entropy, closely related to the idea of uncertainty, stems from the term $\log \frac{1}{p(x)}$ in Equation (2.1). This term is referred to as the *surprisal* for an outcome $x$, and intuitively measures how "surprising" it is to discover that the outcome $x$ has occurred. The surprisal is a monotonically decreasing function of $p(x)$, reflecting the idea that outcomes which occur less frequently are more surprising when they do occur. Also, the surprisal is additive for independent outcomes, so that learning of two independent events together is as surprising as learning of each separately. Consequently, as the expected value of surprisal (Equation (2.1)), entropy is commonly viewed as the average amount of surprise associated with learning the outcome of a variable.

Entropy also extends naturally to two or more variables with the concepts of *conditional* and *joint* entropy. The conditional entropy for discrete variables $X$ and $Y$ is defined by

$$H(X|Y) = \sum_{x} \sum_{y} p(x,y) \log \frac{1}{p(x|y)} \tag{2.2}$$

and measures the residual uncertainty about $X$ when $Y$ is known. Conditional entropy is always less than or equal to the marginal entropy,

$$H(X|Y) \leq H(X), \tag{2.3}$$

so that learning one variable can only ever reduce uncertainty about another. Conditional entropy is zero if, and only if, $X$ is a function of $Y$ (i.e., if for each $x$ there is a $y$ such that $p(x|y) = 1$), in which case there is no uncertainty about $X$ when $Y$ is known. Conversely, $H(X|Y)$ is maximal and equal to $H(X)$ if, and only if, $X$ and $Y$ are independent.

The joint entropy for $X$ and $Y$ is defined by

$$H(X,Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)} \tag{2.4}$$

and is related to the conditional and marginal entropies by the *chain rule for entropy*,

$$H(X,Y) = H(X) + H(Y|X). \tag{2.5}$$

Thus, the uncertainty regarding both $X$ and $Y$ can be decomposed into the uncertainty regarding $X$ plus the residual uncertainty about $Y$ when $X$ is known. From this relationship, it follows immediately that joint entropy is bounded from below by the maximum entropy of any one variable,

$$\max\{H(X), H(Y)\} \leq H(X,Y), \tag{2.6}$$

with equality exactly in the case that $X$ is a function of $Y$, or vice versa. Also, combining the chain rule with Equation (2.3), it follows that joint entropy is *subadditive*,

$$H(X,Y) \leq H(X) + H(Y), \tag{2.7}$$

with equality if, and only if, $X$ and $Y$ are independent. In fact, joint entropy has the even stronger property that it is *submodular* [147, 152, 220, 221], which means that for any two sets of variables $A$ and $B$,

$$H(A \cup B) \leq H(A) + H(B) - H(A \cap B). \tag{2.8}$$

## 2.2 Relative Entropy

The *relative entropy* (also called the *Kullback-Leibler divergence*) for two probability mass functions $p$ and $q$ defined over the same random variable $X$ is given by

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{2.9}$$

In terms of source coding, relative entropy measures the coding inefficiency—that is, the increase in expected code length—that results from assuming that $X$ is distributed according to $q$ when its true distribution is $p$. This can be seen by rewriting the relative entropy as

$$D(p \parallel q) = \left[ \sum_x p(x) \log \frac{1}{q(x)} \right] - \left[ \sum_x p(x) \log \frac{1}{p(x)} \right], \tag{2.10}$$

where the left term is the expected length assuming the distribution is $q$ and the right term is the expected length using the actual distribution $p$.

Relative entropy has the important property that it is always nonnegative,

$$D(p \parallel q) \geq 0, \tag{2.11}$$

with equality if, and only if, the two distributions are identical,

$$p(x) = q(x) \quad \forall x \in X. \tag{2.12}$$

This property, called the *information inequality*, is one of the most fundamental results in all of information theory. For example, it underlies the proof that conditioning always reduces entropy (Equation (2.3)). Because of the information inequality, relative entropy is often thought of as a measure of the distance between two distributions $p$ and $q$. However, it is important to note that relative entropy is not a true distance measure because it is not symmetric and does not satisfy the triangle inequality [36].

## 2.3 Mutual Information

The *mutual information* between two variables $X$ and $Y$ is defined by

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{2.13}$$

and is a measure of the amount of information that $X$ provides about $Y$, and vice versa (note the symmetry with respect to its arguments). This notion of shared information can be grounded in at least two ways: as the *statistical dependence* between $X$ and $Y$, and as the *reduction in uncertainty* that either variable provides about the other. For the former, notice that $I(X;Y)$ can be rewritten as the relative entropy $D(p(x,y) \parallel p(x)p(y))$, where $p(x,y)$ is the true joint distribution and $p(x)p(y)$ is the factorized distribution representing statistical independence between the variables. Thus, $I(X;Y)$ can be thought of as the "distance" from independence between $X$ and $Y$, or equivalently, as the inefficiency of encoding both $X$ and $Y$ while ignoring their dependence. From the information inequality, it follows that $I(X;Y)$ is nonnegative and equals zero if, and only if, $X$ and $Y$ are independent.

Alternatively, $I(X;Y)$ can be written in terms of entropies as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (2.14) \\ &= H(X) + H(Y) - H(X,Y). && (2.15) \end{aligned}$$

In the first equation, $I(X;Y)$ is expressed as the reduction in uncertainty about $X$ when $Y$ is known, while in the second, $I(X;Y)$ measures deviation from the independence bound on joint entropy (Equation (2.7)). It follows that $I(X;Y)$ is bounded from above by the entropy of each individual variable,

$$I(X;Y) \leq \min\{H(X), H(Y)\}, \tag{2.16}$$

with equality exactly in the case that one variable is a function of the other.

The *conditional mutual information* between $X$ and $Y$ given $Z$ is defined as

$$I(X;Y|Z) = \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}, \tag{2.17}$$

and measures the information that $X$ provides about $Y$, and vice versa, when $Z$ is known. Rewriting $I(X;Y|Z)$ as the relative entropy

$$D(p(x,y,z) \parallel p(x|z)p(y|z)p(z)),$$

we see that $I(X;Y|Z)$ measures the distance from conditional independence between $X$ and $Y$ conditioned on $Z$. An alternative expression in terms of entropies,

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z), \tag{2.18}$$

shows that $I(X;Y|Z)$ measures the additional reduction in uncertainty about $X$ that comes from knowing both $Y$ and $Z$ beyond that which comes from knowing $Z$ alone. It is important to note that $I(X;Y|Z)$ can be either greater or less than (or equal to) $I(X;Y)$. Thus, unlike uncertainty, conditioning does not necessarily reduce information. If $I(X;Y|Z)$ is greater than $I(X;Y)$, it is said that $Z$ "enhances" the correlation between $X$ and $Y$. For example, if battery death and fuel blockage are the two possible causes for a car that won't start, then fixing the common effect induces a dependency between the causes: knowing that the car won't start, a healthy battery indicates that the fuel must be blocked [172]. On the other hand, if $I(X;Y|Z)$ is less than $I(X;Y)$, then $Z$ "accounts for" or "explains away" some of the correlation between $X$ and $Y$. For example, since clouds both produce rain and make it dark outside, the correlation between rain and darkness is partly accounted for by the presence of clouds.

Similar to entropy, mutual information obeys a *chain rule for information*,

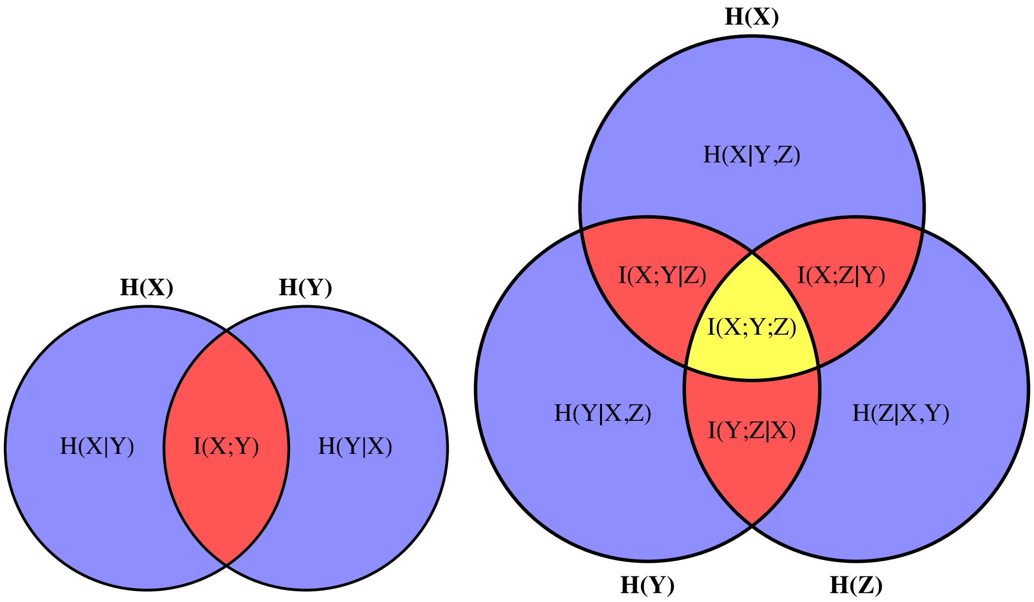
$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z). \tag{2.19}$$

Thus, the total information that $Y$ and $Z$ provide about $X$ can be decomposed into the information that $Z$ provides plus the additional information that $Y$ provides when $Z$ is known. The chain rule for information can also be used to prove an important result called the *data processing inequality*, which states that no amount of processing can ever increase the information that one variable provides about another. In this context, one considers a situation where $X$ and $Z$ are conditionally independent given $Y$, that is, $I(X;Z|Y) = 0$. This situation is also denoted $X \to Y \to Z$, and it is said that $X$, $Y$, and $Z$ form a Markov chain in that order, or that $Y$ screens off $Z$ from $X$. Then the data processing inequality says that $Z$ can provide no more information about $X$ than $Y$ does. In particular, if $Z = g(Y)$, which implies $X \to Y \to Z$, then the inequality says that any function of $Y$, or any processed version of $Y$, can be no more informative than $Y$ itself.

**Theorem 2.1** (data processing inequality)**.** *If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.*

*Proof.* Using the chain rule for information,

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z|Y) \\ &= I(X;Z) + I(X;Y|Z). \end{aligned}$$



(a)

(b)

Figure 2.1: Information diagrams for (a) 2 and (b) 3 variables.

Thus, since $I(X;Z|Y) = 0$, $I(X;Y) \geq I(X;Z)$. □

The relationships between all of the quantities introduced so far can be represented schematically using *information diagrams* (Figure 2.1). In these diagrams, the entropy of each individual variable is represented by a region, and the information shared by any two variables is represented by the overlap of their respective regions. In [258], Yeung showed that these diagrams, which he called I-diagrams, can be formally derived as representations of Shannon information as a measure, in the rigorous sense of measure theory, over sets representing the entropies of individual variables. In terms of these sets, joint entropy, conditional entropy, and mutual information are formally analogous to set union, set difference, and set intersection, respectively. However, unlike standard Venn diagrams, I-diagrams represent *signed* measures, meaning that regions of the diagrams may correspond to either positive or negative values. In particular, the region labeled $I(X;Y;Z)$ in Figure 2.1—representing the three-way intersection of $H(X)$, $H(Y)$, and $H(Z)$—can be either positive or negative. This quantity, called the *interaction information*, and the strange property that it is sometimes negative will be explored at length in Chapter 4.

## 2.4 Specific Information

Mutual information measures the *average* information that one variable provides about another, where the average is taken over all outcomes of the two variables. Thus, by expanding or unrolling that average in various ways, one can explore the specific dependencies between outcomes of the variables that collectively constitute their overall relationship. By unrolling the average with respect to both variables, one can quantify the relationship between individual outcomes $x$ and $y$ using the *pointwise mutual information*,

$$\mathrm{pmi}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}. \tag{2.20}$$

For example, this measure is widely used in computational linguistics to determine the co-occurrence relationships between individual words in a corpus [32,238].

Alternatively, the average mutual information can be unrolled with respect to only one of the two variables in order to quantify the information associated with particular outcomes of that variable. This idea is discussed in [49], where a measure of *specific information* relating an outcome $x$ and a variable $Y$ is defined implicitly by

$$I(X;Y) = \sum_x p(x) I_{\text{spec}}(X = x; Y). \tag{2.21}$$

That is, the specific information $I_{\text{spec}}(X = x; Y)$ defines a class of measures with the only constraint that its expected value with respect to $X$ recovers the mutual information $I(X;Y)$. As pointed out in [49], there are actually infinitely many measures satisfying this constraint, but several candidates stand out as especially attractive choices.

One such candidate is the *response-specific information* [49],

$$I_{\text{rsi}}(R = r; S) = H(S) - H(S|R = r), \tag{2.22}$$

proposed in the context of neural coding to quantify the information that a particular neural response $r$ provides about a stimulus ensemble $S$. The response-specific information quantifies the reduction in uncertainty about the entire stimulus ensemble that results from observing a particular neural response. This measure has the unique property that it is the only definition of specific information that is additive, meaning that

$$I_{\text{rsi}}(R_1 = r_1, R_2 = r_2; S) = I_{\text{rsi}}(R_1 = r_1; S) + I_{\text{rsi}}(R_2 = r_2; S|R_1 = r_1). \tag{2.23}$$

Thus, response-specific information can be decomposed in a manner analogous to the (average) mutual information. However, unlike mutual information, the response-specific information can be either positive or negative, with the latter indicating situations where a particular neural response actually leaves us with more uncertainty about the stimulus.

To distinguish the different role played by stimuli as opposed to responses, Butts [28] proposed an alternative measure called the *stimulus-specific information*:

$$I_{\text{ssi}}(S = s; R) = \sum_r p(r|s) I_r(R = r; S). \tag{2.24}$$

In words, the stimulus-specific information quantifies the informational contribution of a particular stimulus as the weighted average of the response-specific information for all responses associated with that stimulus. Thus, a given stimulus will have a high amount of stimulus-specific information if it tends to evoke responses that are highly informative about the entire stimulus ensemble. Like the response-specific information, this measure can also be either positive or negative.

Finally, both [49] and [28] also discuss the measure of specific information defined by

$$I(S = s; R) = \sum_r p(r|s) \log \frac{p(r|s)}{p(s)}. \tag{2.25}$$

This is by far the most widely used measure of specific information [12, 24, 57, 120, 227], and is the only one that will be used in this thesis. Applying Bayes' rule, $I(S = s; R)$ can be rewritten as

$$I(S = s; R) = \sum_r p(r|s) \left[ \log \frac{1}{p(s)} - \log \frac{1}{p(s|r)} \right]. \tag{2.26}$$

Recalling that $\log \frac{1}{p(s)}$ is the surprisal associated with a particular outcome $s$, we see that $I(S = s; R)$ measures the expected reduction in surprise about $s$ given knowledge of $R$. Thus, in contrast with $I_{\text{ssi}}(S = s; R)$, which weights each response $r$ according to the information that it contributes about the *entire* ensemble $S$, $I(S = s; R)$ quantifies only the information that $R$ provides about the particular outcome $S = s$. The specific information $I(S = s; R)$ has two important properties, which we set apart as lemmas since they will come up again in the technical developments of Chapter 4.

**Lemma 2.1.** *$I(S = s; R) \geq 0$ with equality if, and only if,*

$$p(s, r) = p(s)p(r) \quad \forall r \in R.$$

*Proof.* It follows from the information inequality, since $I(S = s; R) = D(p(r|s) \parallel p(r))$. □

Thus, $I(S = s; R)$ is nonnegative and equals zero exactly in the case that $R$ provides no information about the outcome $S = s$. In fact, it is proven in [49] that $I(S = s; R)$ is the only measure of specific information that is strictly nonnegative.

**Lemma 2.2.** *If $\mathbf{A}$ and $\mathbf{B}$ are sets of random variables with $\mathbf{A} \subseteq \mathbf{B}$, then $I(S = s; \mathbf{A}) \leq I(S = s; \mathbf{B})$.*

*Proof.* Let $\mathbf{C} = \mathbf{B} \setminus \mathbf{A}$. Then we have

$$\begin{aligned}
&I(S = s; \mathbf{B}) - I(S = s; \mathbf{A}) \\
&= \sum_{\mathbf{b}} p(\mathbf{b}|s) \log \frac{p(s, \mathbf{b})}{p(s)p(\mathbf{b})} - \sum_{\mathbf{a}} p(\mathbf{a}|s) \log \frac{p(s, \mathbf{a})}{p(s)p(\mathbf{a})} \\
&= \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{a}, \mathbf{c})}{p(s)p(\mathbf{a}, \mathbf{c})} - \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{a})}{p(s)p(\mathbf{a})} \\
&= \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(\mathbf{c}|\mathbf{a}, s)}{p(\mathbf{c}|\mathbf{a})} \\
&= \sum_{\mathbf{a}} p(\mathbf{a}|s) \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{a}, s) \log \frac{p(\mathbf{c}|\mathbf{a}, s)}{p(\mathbf{c}|\mathbf{a})} \\
&= \sum_{\mathbf{a}} p(\mathbf{a}|s) D(p(\mathbf{c}|\mathbf{a}, s) \parallel p(\mathbf{c}|\mathbf{a})) \geq 0.
\end{aligned}$$

□

Thus, if one set of variables is a superset of another, then the former provides at least as much specific information as the latter. The same is also true of mutual information, which follows from the chain rule for information and the nonnegativity of conditional mutual information.

To illustrate the meaning of $I(S = s; R)$, consider a simplified situation where $S$ is uniformly distributed and $R$ is a function of $S$. For example, in the context of neural coding, this could be thought of as a uniformly sampled stimulus and an idealized neural response with no so-called neuronal noise [24], $H(R|S) = 0$. Then, having observed the response, the only possible source of uncertainty about the stimulus is if the same response is produced by several different stimuli, i.e., if there is *degeneracy* in the mapping from stimuli to responses. To formalize this idea, we define the degree of degeneracy for a particular $\hat{s} \in S$ as

$$\deg(\hat{s}) = |\{s \in S : p(r|s) = p(r|\hat{s})\}|, \tag{2.27}$$

or the number of stimuli that produce the same response as $\hat{s}$ does. Thus, intuitively, $I(S = s; R)$ should be maximal when $\deg(s) = 1$, indicating that $s$ produces a unique response which distinguishes it from all other stimuli, and $I(S = s; R)$ should be minimal when $\deg(s) = |S|$, indicating that all stimuli produce the exact same response. Indeed, this is exactly the behavior of $I(S = s; R)$; using the fact that

$$p(s|r) = \begin{cases} \frac{1}{\deg(s)} & \text{if } p(r|s) = 1 \\ 0 & \text{otherwise,} \end{cases}$$

we can rewrite Equation (2.26) as

$$I(S = s; R) = \log |S| - \log \deg(s), \tag{2.28}$$

so that $I(S = s; R)$ is a monotonically decreasing function of $\deg(s)$. Thus, $I(S = s; R)$ attains its maximum value of $\log |S|$ when $\deg(s) = 1$, in which case $R$ uniquely identifies $s$ out of $|S|$ equally likely possibilities, and $I(S = s; R)$ decreases by one bit every time $\deg(s)$ doubles in size.

## 2.5 Entropy Rate

The remaining measures discussed in this chapter apply to stochastic processes, or time-indexed sequences of random variables. Thus, in contrast with the measures introduced above, all of these measures characterize properties that are essentially dynamic in nature. We will denote a stochastic process by a bi-infinite sequence of random variables

$$\overleftrightarrow{X} = \ldots X_{-1} X_0 X_1 X_2 \ldots \tag{2.29}$$

where $X_i$ represents the state of the process at time $i$. Furthermore, we will assume that all processes are *stationary*, meaning that the joint distribution for any subsequence of variables is time-shift invariant:

$$p(x_0, x_1, \cdots, x_L) = p(x_i, x_{i+1}, \cdots, x_{i+L}) \quad \forall i, L. \tag{2.30}$$

The *entropy rate* for a stochastic process $X$ is defined by

$$h_\mu = \lim_{L \to \infty} \frac{H(X_1, \ldots, X_L)}{L}, \tag{2.31}$$

where this limit is guaranteed to exist for (at least) all stationary processes [36]. From this definition, it is clear that $h_\mu$ corresponds to a kind of density of information content (entropy) per unit time, and for this reason $h_\mu$ is also commonly referred to as the *entropy density*. For stationary processes, the entropy rate can alternatively be expressed in terms of conditional entropy as

$$h_\mu = \lim_{L \to \infty} H(X_L | X_{L-1}, X_{L-2}, \ldots, X_1). \tag{2.32}$$

Thus, the entropy rate quantifies the remaining uncertainty regarding the next state of $X$ when all previous states are known. Consequently, $h_\mu$ is commonly interpreted as the irreducible randomness or unpredictability of $X$, the uncertainty inherent in the next state of $X$ when correlations with all previous states are removed. For example, the entropy rate is zero for periodic processes (which are completely predictable once a full period has been observed) and even for processes with infinite memory, so long as they are deterministic and thus are perfectly correlated over some timescale [40]. In contrast, the entropy rate is positive for processes with some amount of intrinsic stochasticity, such as Markov processes or independent and identically distributed (IID) processes. For instance, the entropy rate for a sequence of fair coin flips is one bit, since knowing the outcomes of previous flips does nothing to reduce uncertainty regarding the next flip.

## 2.6 Excess Entropy

Closely related to the entropy rate is a measure called the *excess entropy* [22, 40, 55], which quantifies the amount of structure or memory in a stochastic process. The excess entropy for a process $X$ is defined by

$$\mathbf{E} = \sum_{L=1}^{\infty} \Big[ h_\mu(L) - h_\mu \Big], \tag{2.33}$$

where

$$h_\mu(L) = \frac{H(X_1, \ldots, X_L)}{L} \tag{2.34}$$

is the length-$L$ approximation of the entropy rate. Each of these length-L approximations overestimate the actual entropy rate, due to the fact that temporal correlations over timescales longer than $L$ are missed by $h_\mu(L)$ but picked up by $h_\mu$. Thus, each overestimate $[h_\mu(L) - h_\mu]$ quantifies the apparent uncertainty at scale $L$ that is accounted for by longer-range correlations, and, by summing these overestimates over all $L$, the excess entropy measures the total amount of correlation or structure over all timescales in a process. This interpretation becomes more apparent when the excess entropy is expressed as a mutual information,

$$\mathbf{E} = I(\overleftarrow{X}; \overrightarrow{X}), \tag{2.35}$$

where $\overleftarrow{X} = \ldots X_{-3} X_{-2} X_{-1}$ and $\overrightarrow{X} = X_0 X_1 X_2 \ldots$ represent the semi-infinite past and future of $X$, respectively. Thus, the excess entropy can alternatively be thought of as the total amount of information that the past of $X$ provides about its future, or equivalently, how well future states of $X$ can be predicted from its past. For this reason, excess entropy is also commonly referred to as the *predictive information* [180]. In applications of excess entropy, as well as of entropy rate, one is often interested not only in their limiting values but also in the rate of convergence towards those values, an idea that will be discussed further in the next chapter.

## 2.7 Transfer Entropy

The final concept introduced in this chapter is the *transfer entropy* [113, 198], which provides a general measure of the influence that one stochastic process has on another. The transfer entropy from one process $Y$ to another process $X$ is defined by

$$T_{Y\to X} = I(X_t; Y_{t-1}^{(j)} | X_{t-1}^{(k)}), \tag{2.36}$$

where $Y_{t-1}^{(j)} = \{Y_{t-1}, \ldots, Y_{t-j}\}$ represents the $j$ previous states of $Y$, and likewise for $X_{t-1}^{(k)}$ and $X$[3]. Thus, transfer entropy quantifies the information that previous states of $Y$ provide about the next state of $X$ when conditioned on $X$'s own history. The idea behind this definition is that conditioning on the previous states of $X$ removes the information shared by $X$ and $Y$ due to common histories or inputs, thereby isolating the information that is transferred from $Y$ to $X$. In particular, transfer entropy was originally proposed as an alternative to the time-delayed mutual information (TDMI) $I(X_t; Y_{t-1})$—which was the standard measure of information transfer prior to $T_{Y\to X}$ [114, 239]—for the reason that TDMI confounds information transfer with shared information [198].

Two other interpretations of $T_{Y\to X}$ are also worth noting, as they serve to connect it with concepts introduced above. First, by expressing $T_{Y\to X}$ in terms of relative entropies,

$$T_{Y\to X} = \sum_{x_{t-1}} p(x_{t-1}) D(p(x_t|x_{t-1}, y_{t-1}) \parallel p(x_t|x_{t-1})), \tag{2.37}$$

[3] Except where otherwise noted, we will henceforth assume the standard setting of $j = k = 1$ and omit the superscripts $(j)$ and $(k)$.

we see that $T_{Y\to X}$ is zero if, and only if, the *generalized Markov property* holds:

$$p(x_t|x_{t-1},y_{t-1}) = p(x_t|x_{t-1}) \quad \forall x_t, x_{t-1}, y_{t-1}. \tag{2.38}$$

Thus, $T_{Y\to X}$ is zero exactly in the case that $Y$ has no influence on the transition probabilities of $X$, with non-zero values indicating the magnitude of this influence. In other words, $T_{Y\to X}$ can be interpreted as measuring deviation from the generalized Markov property. Second, taking the limits $j \to \infty$ and $k \to \infty$ and rewriting $T_{Y\to X}$ in terms of entropies,

$$\begin{aligned} T_{Y\to X} &= H(X_t|X_{t-1}^{(\infty)}) - H(X_t|X_{t-1}^{(\infty)}, Y_{t-1}^{(\infty)}) \\ &= h_\mu - H(X_t|X_{t-1}^{(\infty)}, Y_{t-1}^{(\infty)}), \end{aligned} \tag{2.39}$$

transfer entropy can be seen as a generalization of the entropy rate to more than one process, effectively yielding a mutual information rate. In this context, $T_{Y\to X}$ measures the additional reduction in uncertainty that the infinite histories of both $Y$ and $X$ provide about the next state of $X$, beyond that provided by the infinite history of $X$ alone.

## 2.8 Correlation, not Causation

As a concluding note, we wish to emphasize that all of the measures introduced here are fundamentally measures of correlation, and thus cannot be used directly to infer causal interactions. In particular, all of the measures are defined for *observational* probability distributions—i.e., distributions that reflect a system's uninterrupted behavior—whereas causal inference requires the use of *interventional* or *perturbational* methods, in which certain variables are disrupted and the influence on other variables is observed [10,171]. Thus, although informational measures can yield insights into causal interactions, the standard proviso that correlation does not equal causation should be kept in mind throughout.

# 3

# Overview of Related Work

The framework developed in this thesis has four central and distinguishing characteristics. First, at its core is a novel method for decomposing *multivariate* information, allowing us to consider more complex informational relationships than are possible with standard univariate and bivariate techniques. Second, our framework is designed to capture *dynamic* properties of information flow, in contrast with the traditional focus of information-theoretic techniques on static information structure. Third, our framework is based on measures of *specific* information, which permit a more fine-grained level of analysis than the measures of average information that are more commonly used. Fourth, our framework is designed for the analysis of *embodied* systems, enabling us to explore the specific ways that an agent's body and environment can influence the dynamics of information flow. Each of these key features—summarized by the terms *multivariate*, *dynamic*, *specific*, and *embodied*—appears in existing strands of research, though our framework is unique in incorporating all of them.

In this chapter, we discuss examples of recent work applying information theory in neuroscience, complex systems, and embodied cognitive science that incorporate one or more of the key features of our framework. These applications also by and large represent the cutting-edge of work in theoretical and applied information theory, so they provide a natural context for framing and evaluating the novel contributions of this thesis. In the four sections of this chapter, we review examples of (1) multivariate, (2) dynamic, (3) specific, and (4) embodied information-theoretic analyses. Later, in Chapter 8, we revisit these four key features in the context of how our framework relates to, unifies and extends this existing work.

## 3.1 Multivariate Information

Perhaps the most prolific experimental application of information theory has been in the area of neural coding [45,187]. Research in this area is primarily concerned with three interrelated questions about how neural responses carry information about stimuli. First, what stimulus features are encoded, and with what precision? Second, what aspects of the neural response are doing the encoding? Third, how efficient is the encoding? Historically, researchers have used information theory to explore these questions by quantifying the entropy of individual neurons and the mutual information between neural responses and stimulus ensembles [110,185,243]. For instance, the precision with which a stimulus ensemble $S$ is encoded by a neural response $R$ can be determined by comparing the mutual information $I(S;R)$ with the entropy $H(S)$, where equality between these two quantities implies that the stimulus is encoded with perfect precision [48]. Likewise, questions of encoding efficiency can be addressed by comparing $I(S;R)$ with $H(R)$, where the latter represents the maximum information capacity for the neural response [59,223].

More recently, and spurred by advances in multielectrode recording techniques [27], improved methods for estimating information-theoretic quantities [160,168,193,243], and greater computational resources, the focus of research in neural coding has shifted towards the use of multivariate informational measures. In particular, numerous studies have explored whether the information encoded by different neurons, or by the same neuron across time, is synergistic, redundant, or independent [25,69,131,158,169,196,225]. As conventionally understood, synergy, redundancy, and independence represent mutually exclusive alternatives for how information may be encoded, with the difference between them boiling down to whether or not information is additive[1]. In particular, synergy corresponds to situations where the information that neural responses $R_1$ and $R_2$ provide is superadditive,

$$I(S; R_1, R_2) > I(S; R_1) + I(S; R_2),$$

whereas redundancy corresponds to subadditivity (reversing the inequality) and independence corresponds to additivity (replacing the inequality with an equality). Thus, for instance, if $R_1$ and $R_2$ represent the responses of different neurons, then synergy indicates that their joint response provides information that is not available from either neuron individually, indicating some form of population coding [7, 131, 169, 185, 196]. In contrast, redundancy suggests that some of the same information is simultaneously carried by both neurons, a property that may be advantageous in providing robustness to the damage of individual neurons [158,182,186,225]. Alternatively, if $R_1$ and $R_2$ represent two different features of a single neuron's response, e.g., the occurrence of spikes at two different times, then synergy indicates that compound events in the spike train carry information beyond that carried by the individual parts, which sheds light on the question of what aspects of the neural response are doing the encoding [25].

Multivariate informational measures have also figured prominently in the field of complex systems, a fact that is unsurprising given the field's interest in systems with many interacting components. In particular, multivariate measures have been used to define and quantify the concept of complexity itself, which has been a topic of intense interest and debate since the field's inception [9, 61, 141, 180]. One particularly influential example is the measure of *neural complexity* introduced by Tononi, Sporns and Edelman [230,232,233].

[1] In contrast, a key property of PI-decomposition, which will be introduced in Chapter 4, is that synergistic, redundant, and independent (or, as we will call them, unique) informational contributions are *not* mutually exclusive, but rather can occur simultaneously. Indeed, as we will see, this is a crucial feature distinguishing our approach from previous methods of quantifying multivariate information.

The neural complexity for a set **X** of $N$ random variables $\{X_1, X_2, \ldots, X_N\}$, which might correspond to a collection of neural units, is defined as

$$C_N(\mathbf{X}) = \sum_{k=1}^{N} [\frac{k}{N} I(\mathbf{X}) - \langle I(\mathbf{X}_j^k) \rangle] \tag{3.1}$$

where

$$I(\mathbf{X}) = \sum_{i=1}^{N} H(X_i) - H(\mathbf{X}) \tag{3.2}$$

denotes the *integration*[2] [231] for the elements in **X** and $\langle I(\mathbf{X}_j^k) \rangle$ denotes the average integration for all subsets of size $k$ in **X**. This measure is designed to capture the intuition that complex systems combine the properties of high local segregation and global integration. For example, the brains of higher vertebrates, which may be thought of as prototypical examples of complex systems, combine high functional segregation in individual brain regions with high global cooperation in the production and coordination of behavior.

Neural complexity (Equation (3.1)) captures these dual properties by considering the integration across all spatial scales in a system (i.e., across all subsets of units). If there is no local segregation at different spatial scales (i.e., different subset sizes), then the average integration will grow linearly as a function of subset size, and for subsets of size $k$ will equal $\frac{k}{N} I(\mathbf{X})$. Note that this corresponds to the first term inside the summation in Equation (3.1). If, on the other hand, there is a high amount of local segregation at different spatial scales, then the average integration for subsets of size $k$ will be less than $\frac{k}{N} I(\mathbf{X})$, contributing to an increase in neural complexity. Thus, in general, neural complexity will be high if there is a high level of global integration and a high amount of local segregation at different spatial scales. Recent work has shown that neural complexity can also be tied to specific topological features of networks [13, 14, 47, 217, 257], and that it is systematically related to a number of other information-theoretic complexity measures [9]. In addition, neural

[2] Integration is also called total correlation [247], among other names, and will be discussed further in Section 4.1.

complexity has been applied in several studies of embodied cognitive systems, which we will discuss in Section 3.4.

## 3.2 Dynamic Information

In neuroscience, dynamic informational measures have increasingly been used for the identification and analysis of functional brain networks. A functional brain network refers to a network of brain regions that are related by statistical dependencies and patterns of information flow, as opposed to structural or anatomical networks, which are determined by anatomical connections between brain regions [64, 213]. While functional networks are undoubtedly related to and constrained by their structural counterparts, they differ in that they are highly dynamic and variable, sometimes rapidly reconfiguring in response to external inputs or changing task demands. Thus, the identification of functional brain networks provides insight into how brain regions interact during specific tasks or behaviors, as well as how quickly and in what ways these interactions change over time.

The dynamic informational measures that have been used most widely to this end are the transfer entropy (Section 2.7, see also [241]) and Granger causality [53,83]—essentially a linear version of transfer entropy [3,13]—though other informational measures have also been proposed [4, 97, 173, 183, 209]. For example, one recent study used transfer entropy to identify functional networks over a range of timescales in a large-scale simulation of the macaque neocortex, and then explored the relationship between these functional networks and the underlying structural connectivity [98]. The authors found that over long timescales the functional networks largely aligned with the underlying structural topology, but at intermediate and fast timescales the networks exhibited a range of interesting dynamical behaviors, including rapid synchronization and desynchronization of brain regions and a large number of metastable states. More recently, transfer entropy has been used to explore interactions between brain regions during a visuomotor tracking task, and to examine how these interactions varied with changing task difficulty [136]. Task difficulty was found to modulate the amount of involvement of specific brain regions, with regions involved in movement planning and fine motor control playing more active roles as task difficulty increased. In other studies, dynamic informational measures have also been used to explore functional interactions in the human visual cortex [96], cat auditory cortex [80], simulated neocortical columns [161], and between brain areas of macaques [133].

In complex systems, another dynamic informational measure that has received considerable attention is the excess entropy (Section 2.6), which has been advocated by several authors as an alternative measure of complexity [22, 40, 61, 84, 132, 135, 180]. That is, in contrast with the measure of neural complexity discussed in the previous section, which essentially defines complexity as a static property of differentiated structure over a range of spatial scales, excess entropy measures complexity in terms of a system's behavior over time[3]. The basic idea is that excess entropy reflects the amount of "memory" that a system contains, or the timescale over which correlations persist in the system's behavior. Thus, systems with larger amounts of memory, or whose future behavior depends on events that occurred in the more distant past, have greater complexity. Equivalently, excess entropy can be thought of as the amount of information that one would need to know about a system's past in order to make the best possible prediction about its future [180].

Most impressively, it has been shown that the rate at which excess entropy grows over time can be used to assign systems to one of several complexity classes [22, 23, 40, 180]. In general, excess entropy is always nonnegative and grows with time less rapidly than

[3] Although we will focus on this dynamic interpretation here, excess entropy can also be interpreted as a measure of spatial complexity. The only difference is that the semi-infinite 'past' and 'future' of a random process, as appear in Equation (2.35), are instead taken to be the semi-infinite 'left' and 'right' halves, respectively, of a spatially-extended system. For example, with this interpretation excess entropy has been used to quantify the complexity of one-dimensional spin systems [39].

a linear function (i.e., it is subextensive). Thus, in the simplest case that the excess entropy remains finite, then regardless of how long a system is observed there is only a finite amount of information to be gained about its future. For example, it is possible to completely predict the behavior of a periodic regular process once one full period has been observed. This corresponds to the simplest possible complexity class. Alternatively, if the excess entropy diverges, indicating that the future of a system is influenced by events in the arbitrarily distant past, then the growth rate may be either slow (logarithmic) or fast (sublinear power). The former case indicates that a system can be modeled with a finite number of parameters, where the number of parameters is given by the coefficient of this divergence. The latter case corresponds to systems that require an infinite number of parameters to model, representing the highest level of complexity.

## 3.3 Specific Information

Compared with the other key features of our framework, that of specific information is certainly the least explored. Indeed, the only relevant applications that we are aware of are those in neural coding that were mentioned in Section 2.4. As alluded to in Section 2.4, there are two ways in which measures of specific information have been applied to study neural codes. First, they have been used to quantify the informativeness of particular neural responses with respect to a stimulus ensemble [49]. Second, they have been used to determine how well particular stimuli are encoded by neural responses [28]. The latter is by far the more widely explored area, and is akin to how specific information will be used in our framework. As described in [24], using specific information in this way allows one to replace the traditional stimulus-response curves used in studies of neural coding with stimulus-information curves—i.e., plots of the specific information $I(S = s; R)$ as a function of stimulus condition $s$—where the latter curves convey how well an ideal observer could discriminate between the stimulus conditions by observing the neural response. For example, this kind of analysis was used early on to analyze microelectrode recordings from the afferent visual system of the cat [56,57], and more recently has been used to characterize the encoding of information in the cercal system of crickets [227] and the encoding of spatial pattern information in the primary visual cortex of monkeys [120].

## 3.4 Information in Embodied Systems

Information-theoretic techniques have been used in several recent studies to analyze the flow of information in embodied agents. The central idea to emerge from this research is the principle of *information self-structuring*, which is a specific idea about the importance of embodiment for cognition [144,159,174–176,179,195,216]. Information self-structuring is the idea that, beyond imposing physical and energetic constraints on an agent's behavior, the importance of having a body is that it allows an agent to actively select and structure the information that it receives from its environment. In other words, rather than passively receiving information in the form of sensory stimulation, an embodied agent can actively shape the information that it receives through a continual process of sensorimotor coordination.

The principle of information self-structuring has been demonstrated in several studies, each having the same general form [143–145]. These studies employ a robotic active vision system consisting of a color camera mounted on the end of a robotic "arm". This vision system is studied in two different behavioral modes, corresponding to different experimental conditions. In the first behavioral mode, the vision system is programmed to actively locate and track regions of red inputs in its environment. Thus, in this case the system's behavior is governed by a continual process of sensorimotor coordination. In the second behavioral mode, the visual input and motor signals are decoupled, so that the system exhibits no sensorimotor coordination. The basic idea, then, is to use information-theoretic techniques to compare the coordinated and uncoordinated sensorimotor activity, with the hypothesis being that coordinated behavior will result in greater informational structure.

In analyzing the robotic vision system, entropy and mutual information were first used to characterize the red channel of the visual inputs received by the camera, since the system was configured to track red objects in the tracking condition. The main results were a significantly lower entropy for central pixels in the tracking condition, as compared to peripheral pixels in the tracking condition and all pixels in the uncoordinated condition, and a significantly higher mutual information between pairs of central pixels in the tracking condition, as compared to all other pairs of pixels. Intuitively, these results make sense since the camera in the tracking condition tends to position red objects in the center of its visual field, meaning that the red values for these pixels will tend to be consistently high (reducing entropy) and similar to one another (increasing mutual information). In terms of information self-structuring, these results also reflect a systematic difference in the informational structure produced by coordinated sensorimotor activity: coordination results in lower entropy and higher mutual information for a certain region of visual inputs.

In further studies of the robotic vision system [145, 214], transfer entropy was used to quantify the amount of information exchanged between various sensor, neural, and motor components. Here the authors found significantly more directed information transfer in the coordinated tracking condition as compared with the uncoordinated condition. This, of course, is not surprising, since one would expect coordination to require structured causal interactions, such that changing sensor inputs will drive compensatory changes in motor control. Another experiment, performed on the robot in the tracking condition, varied how quickly red objects were displaced from the center of the visual field once the robot had achieved fixation. When the red object was displace slowly or remained still, only small amounts of information transfer were detected. On the other hand, when the object made significant "jumps" away from the center of the visual field, there was a significant increase in the strength of directed interactions. When the object moves away abruptly, the vision system must make rapid compensatory movements in order to track it, and transfer entropy was found to reflect these changing motor patterns. Other studies have also explored how patterns of information transfer vary as a result of morphological changes and learning [145].

Finally, the measures of integration and complexity discussed in Section 3.1 were also applied to the robotic vision system and were found to be higher in the tracking condition relative to the uncoordinated condition. Several other studies have also reported a general increase in both integration and complexity in cases of coordinated versus uncoordinated behavior [203,255,256]. A related complexity measure, called causal density [204], has also been shown to exhibit a general increase with the difficulty of behavioral coordination in several simulated and robotic agents [128,201,202]. In an interesting twist, it has also been shown that evolving agents with a fitness function based on integration or complexity can produce agents with a high level of behavioral coordination[4] [215]. Thus, together these studies demonstrate a sort of bidirectional relationship between the measures of integration and complexity and the presence of coordinated behavior.

[4] In fact, this study only scratches the surface of the growing research area in *guided self-organization*, which explores the use of informational measures as fitness functions for evolving adaptive behavior. A variety of informational measures—including transfer entropy [137] and predictive information [8]—have been explored to this end, with the upshot being that intelligent solutions often result when agents are selected simply to maximize informational structure [111,122–125,181,261].

# 4

# Decomposing Multivariate Information

This chapter develops the theoretical foundation for our approach to information dynamics. The central contribution described here is a novel method for decomposing the Shannon information in a multivariate system, called *partial information decomposition*. Such a method is essential for our purposes because the quantities of interest for information dynamics are fundamentally multivariate in nature. For example, in the simplest case of information flow for a single component, one must consider at least three variables: one representing the stimulus and two representing the states of the component at different times. Meanwhile, there are several known limitations and inconsistencies with existing definitions of multivariate information.

The chapter begins with a discussion of previous attempts to extend mutual information to multivariate interactions. Then the method of partial information decomposition is developed, beginning from a first principles analysis of the structure of multivariate information. This is followed by a discussion of how partial information decomposition relates to interaction information, the current de facto measure of multivariate information. Finally, the chapter concludes with a brief sketch of directions for future work.

## 4.1 Multivariate Extensions of Mutual Information

Perhaps the most widely used concept from information theory is Shannon's mutual information, which, as described in Section 2.3, provides a general measure of the interdependence between two variables, or two sets of variables. Given this general applicability, there has also been great interest in extending mutual information to multivariate interactions, but despite numerous attempts [2,41,79,92,153,197,224,247] this remains largely an open problem. Meanwhile, many of the most interesting and challenging scientific questions, such as many-body problems in physics, $n$-person games in game theory, and population coding in neuroscience, involve understanding the structure of interactions between three or more variables.

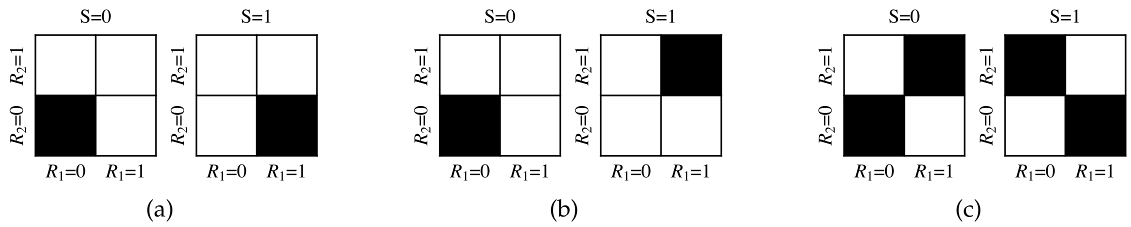
The two main attempts to generalize mutual information to multivariate interactions are the *total correlation* introduced by Watanabe [247] (also known as the multivariate constraint [66], multiinformation [224], and integration [232]) and the *interaction information* of McGill [153] (also known as multiple mutual information [92], co-information [21], and synergy [69]). The total correlation, as its name suggests, measures the total amount of dependency between a set of variables as a single monolithic quantity. Consequently, the total correlation does not provide any insight into how dependencies are distributed amongst the variables, i.e., it says nothing about the *structure* of multivariate information.

In contrast, interaction information was proposed as a measure of the amount of information bound up in a set of variables beyond that which is present in any subset of those variables. Thus, entropy and mutual information correspond to first- and second-order interaction information, respectively, and together with its third-, fourth-, and higher-order variants, interaction information provides a natural way of characterizing the structure of multivariate information. As mentioned in Section 2.3, interaction information is also the natural generalization of mutual information when Shannon entropy is viewed as a signed measure on information diagrams [21, 36, 259]. However, the wider use of interaction information has largely been hampered by the "odd" [21] and "unfortunate" [36] property that, for three or more variables, the interaction information can be negative (see also [92, 226, 236, 259, 262]). For information as it is commonly understood, it is entirely unclear what it means for one variable to provide "negative information" about another. Moreover, as we will show, this confusing property of negativity is symptomatic of deeper problems regarding the interpretation of interaction information for higher-dimensional systems.

In this chapter, we consider the general problem of quantifying multivariate information in a way that illuminates its compositional structure. Thus, the work presented here is in much the same spirit as McGill's seminal analysis of multivariate interactions. However, unlike McGill's interaction information, the approach we take leads to a decomposition of multivariate information into a family of nonnegative partial information measures, each of which supports a clear interpretation as an informational quantity. Moreover, we show that these partial information measures, each of which is atomic in nature, together form a lattice which clarifies the structure of multivariate information, in much the way that set- and lattice-theoretic formulations of the Shannon entropies have helped to clarify standard informational quantities [21, 43, 87, 91, 258, 259]. Finally, our analysis will also show how the confusing negativity of interaction information can be explained by its confounding of redundant and synergistic interactions.

## 4.2 The Structure of Multivariate Information

Suppose we are given a random variable $S$ and a random vector $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$. Then our goal is to decompose the total information that $\mathbf{R}$ provides about $S$ in terms of the partial information contributed either individually or jointly by various subsets of $\mathbf{R}$. For



(a) (b) (c)

Figure 4.1: Examples of (a) unique information, (b) redundancy, and (c) synergy. Each set of tiled squares represents a probability distribution for $S, R_1, R_2 \in \{0, 1\}$. Black tiles represent equiprobable outcomes and white tiles represent zero-probability outcomes.

example, in a neuroscience context, $S$ may correspond to a stimulus that takes on different values and $\mathbf{R}$ to the evoked responses of different neurons. In this case, we would like to quantify the information that the joint neural response provides about the stimulus, and to distinguish between information due to responses of individual neurons versus combinations of them [69,187].

As a preliminary step towards characterizing multivariate information in its full generality, we begin by examining some of its basic properties in the simplest case of a system with three variables, $S$ and $\mathbf{R} = \{R_1, R_2\}$. Intuitively, how should the total information $I(S; R_1, R_2)$ decompose into partial information contributions? In other words, what are the various ways in which subsets of $\mathbf{R}$ might contribute information about $S$? To illustrate the range of possibilities, consider the three examples in Figure 4.1. In each example, there is initially one bit of uncertainty about $S$ and zero bits of uncertainty when both $R_1$ and $R_2$ are known: $H(S) = I(S; R_1, R_2) = 1$. In the first example, Figure 4.1(a), the value of $S$ can be uniquely determined from $R_1$ while $R_2$ provides no information, reflected in the fact that $I(S; R_1) = 1$ and $I(S; R_2) = 0$. We refer to this kind of contribution, where $R_1$ provides information that $R_2$ does not, or vice versa, as *unique information* from $R_1$ or $R_2$, respectively. In this case, the total information from $\mathbf{R}$ reduces to the unique information from $R_1$ and the system actually simplifies to two variables, though in general it is possible for $R_1$ and $R_2$ to provide different unique information. Alternatively, $R_1$ and

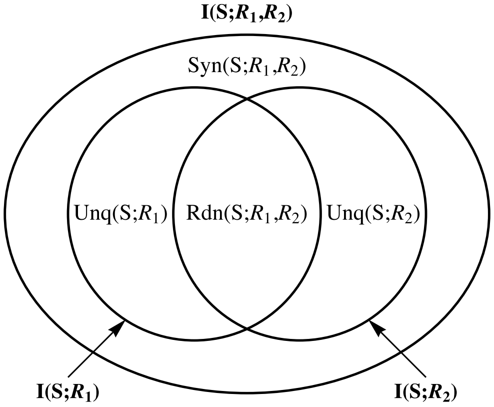


Figure 4.2: Structure of information for 3 variables. Inner regions correspond to unique information (*Unq*), redundancy (*Rdn*), and synergy (*Syn*).

$R_2$ may provide the same or overlapping information, which we call *redundancy*. For example, in Figure 4.1(b), the value of $S$ can be resolved from knowledge of either $R_1$ or $R_2$, $I(S; R_1, R_2) = I(S; R_1) = I(S; R_2)$, so that $R_1$ and $R_2$ redundantly provide complete information. Finally, a third possibility is that the combination of $R_1$ and $R_2$ may provide information that is not available from either alone, which we call *synergy*. A well-known example for binary variables is the exclusive-OR function $S = R_1 \oplus R_2$ (Figure 4.1(c)), in which case $R_1$ and $R_2$ individually provide no information, $I(S; R_1) = I(S; R_2) = 0$, but together provide complete information, $I(S; R_1, R_2) = 1$.

Some reflection should convince the reader that these three possibilities exhaust the alternatives for three variables. That is, the total information provided by $\mathbf{R}$ must come either uniquely from $R_1$ or $R_2$, redundantly from $R_1$ and $R_2$, or synergistically from the combination of $R_1$ and $R_2$. The relationships between these possibilities are depicted schematically in Figure 4.2. The total information provided by $\mathbf{R}$ includes, but is not limited to, the information provided by $R_1$ and $R_2$; thus, $I(S; R_1, R_2)$ is depicted as a superset of $I(S; R_1)$ and $I(S; R_2)$ but not coextensive with $I(S; R_1) \cup I(S; R_2)$. The difference between $I(S; R_1, R_2)$ and $I(S; R_1) \cup I(S; R_2)$ corresponds to synergy. Furthermore, the information provided by $R_1$ and $R_2$ can either overlap, corresponding to redundancy, or not, corresponding to unique information.

In summary, for three variables we can identify unique information, redundancy, and synergy as the basic atoms of multivariate information. In fact, as later developments will clarify, unique information is best thought of as a degenerate form of redundancy or synergy, so that redundancy and synergy alone constitute the building blocks of multivariate information. In particular, we will find that combinations of redundancy and synergy, which may at first sound paradoxical, in fact play a fundamental role in structuring multivariate information in higher dimensions. Next we proceed to formalize these ideas, beginning with the notion of redundancy.

## 4.3 Redundancy as Overlapping Information

Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ be potentially overlapping subsets of $\mathbf{R}$, called *sources*. How can we quantify the redundant information that all sources provide about $S$? The intuition that we want to capture is that redundancy corresponds to overlapping or shared information. In particular, redundancy should behave similarly to the intersection operator from set theory, measuring the common information "contained" by two or more sources.

To capture this idea of intersecting information, we define redundancy as a $(k+1)$-argument function

$$I_\cap : S \times \underbrace{2^{\mathbf{R}} \times \cdots \times 2^{\mathbf{R}}}_{k} \mapsto \mathbb{R}, \qquad (S, \mathbf{A}_1, \ldots, \mathbf{A}_k) \mapsto I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_k) \tag{4.1}$$

that satisfies the following axioms:

**Axiom 4.1** (symmetry). *$I_\cap$ is symmetric in the $\mathbf{A}_i$'s.*

**Axiom 4.2** (self-redundancy). *$I_\cap(S; \mathbf{A}) = I(S; \mathbf{A})$.*

**Axiom 4.3** (monotonicity). $I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_{k-1}, \mathbf{A}_k) \leq I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_{k-1})$ *with equality if* $\mathbf{A}_{k-1} \subseteq \mathbf{A}_k$.

Axiom 4.1 states that the redundancy for a collection of sources should not depend on the order in which they are considered. Axiom 4.2 states that the redundancy for a single source considered by itself, referred to as the 'self-redundancy', is simply equal to the information provided by that source. Finally, Axiom 4.3 states that redundancy can only decrease or remain the same as more sources are added, with redundancy remaining the same if the new source is a superset of an existing one. Note that these axioms are analogous to basic properties of set intersection; namely, that it is commutative (symmetric): $X \cap Y = Y \cap X$; idempotent ('self-intersection' is preserving): $X \cap X = X$; and monotonic: $(X_1 \cap \cdots \cap X_{k-1} \cap X_k) \subseteq (X_1 \cap \cdots \cap X_{k-1})$ with equality if $X_{k-1} \subseteq X_k$. Thus, we claim that any measure of redundancy as intersecting information must satisfy these basic axioms.

Two useful properties of redundancy follow immediately from these axioms.

**Corollary 4.1.** *$I_\cap$ is nonnegative.*

*Proof.* Since $\emptyset \subseteq \mathbf{A}_i$ for any $\mathbf{A}_i$,

$$I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_k, \emptyset) = I_\cap(S; \emptyset) = I(S; \emptyset) = 0$$

by Axioms 4.2 and 4.3. Also from Axiom 4.3, we have that

$$I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_k) \geq I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_k, \emptyset),$$

so we conclude that $I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_k) \geq 0$. □

**Corollary 4.2.** $I(S; \mathbf{A}_i) = I_\cap(S; \mathbf{A}_i) \geq I_\cap(S; \mathbf{A}_i, \ldots)$.

Thus, redundancy is nonnegative, bounded from above by the information from any one source, and maximized by the self-redundancy, analogous to the property that mutual information is maximized by the self-information $I(S;S) = H(S)$.

An appealing candidate measure of redundancy that satisfies the axiomatic definition $I_\cap$ is

$$I_{\min}(S; \mathbf{A}_1, \ldots, \mathbf{A}_k) = \sum_s p(s) \min_{\mathbf{A}_i} I(S = s; \mathbf{A}_i) \tag{4.2}$$

where $I(S = s; \mathbf{A})$ is the specific information that $\mathbf{A}$ provides about each outcome $s \in S$ (Equation (2.26); see Section 2.4 for a discussion of specific information). Thus, $I_{\min}$ quantifies redundancy as the minimum information that any source provides about each outcome of $S$, averaged over all possible outcomes. This definition captures the idea that redundancy is the information shared by all sources (the minimum that any one provides), while taking into account that sources may provide information about different outcomes of $S$.

**Theorem 4.1.** *$I_{\min}$ satisfies the axiomatic definition $I_\cap$.*

*Proof.* That $I_{\min}$ satisfies Axioms 4.1 and 4.2 follows from basic properties of $\min$, while Axiom 4.3 follows from Lemma 2.2. □

Henceforth, we are careful to use $I_\cap$ when discussing properties that follow from the axiomatic definition, and $I_{\min}$ when discussing properties specific to that measure. $I_{\min}$ has the following additional desirable properties.

**Theorem 4.2.** *$I_{\min}(S; \mathbf{A}_1, \ldots, \mathbf{A}_k) = 0$ if, and only if, for every joint outcome $(s, \mathbf{a}_1, \ldots, \mathbf{a}_k)$,*

$$p(s, \mathbf{a}_i) = p(s)p(\mathbf{a}_i)$$

*for at least one $\mathbf{a}_i$.*

*Proof.* It follows trivially from Lemma 2.1 and basic properties of $\min$. □

Theorem 4.2 is a natural generalization of the property that $I(S; \mathbf{A}) = 0$ if, and only if, for every joint outcome $(s, \mathbf{a})$,

$$p(s, \mathbf{a}) = p(s)p(\mathbf{a}).$$

Thus, according to this theorem, redundancy (defined as $I_{\text{min}}$) is zero if, and only if, for each joint outcome one or more of the sources provide no information about $S$.

**Theorem 4.3.** *If* $S \to \mathbf{A}_k \to \mathbf{A}_{k-1}$, *then*

$$I_{\text{min}}(S; \mathbf{A}_1, \ldots, \mathbf{A}_{k-1}, \mathbf{A}_k) = I_{\text{min}}(S; \mathbf{A}_1, \ldots, \mathbf{A}_{k-1}).$$

*Proof.* It suffices to show that if $S \to \mathbf{A}_k \to \mathbf{A}_{k-1}$, then $I(S = s; \mathbf{A}_{k-1}) \leq I(S = s; \mathbf{A}_k)$ for all $s \in S$. If $S \to \mathbf{A}_k \to \mathbf{A}_{k-1}$, then $I(S; \mathbf{A}_{k-1}|\mathbf{A}_k) = 0$ and thus $I(S; \mathbf{A}_{k-1}, \mathbf{A}_k) = I(S; \mathbf{A}_k)$. From Lemma 2.2, this means that $I(S = s; \mathbf{A}_{k-1}, \mathbf{A}_k) = I(S = s; \mathbf{A}_k)$ for all $s \in S$. Also from Lemma 2.2, we have that

$$I(S = s; \mathbf{A}_{k-1}, \mathbf{A}_k) \geq I(S = s; \mathbf{A}_{k-1}),$$

so we conclude that $I(S = s; \mathbf{A}_{k-1}) \leq I(S = s; \mathbf{A}_k)$ for all $s \in S$. □

In words, Theorem 4.3 states that if $S$, $\mathbf{A}_k$, and $\mathbf{A}_{k-1}$ form a Markov chain in that order (i.e., if $\mathbf{A}_k$ screens off $\mathbf{A}_{k-1}$ from $S$), then any information provided by $\mathbf{A}_{k-1}$ is also provided redundantly by $\mathbf{A}_k$. As the following corollary shows, this theorem is a generalization of the data processing inequality discussed in Section 2.3.

**Corollary 4.3** (data processing inequality). *If* $S \to \mathbf{A}_i \to \mathbf{A}_j$, *then* $I(S; \mathbf{A}_i) \geq I(S; \mathbf{A}_j)$.

*Proof.* If $S \to \mathbf{A}_i \to \mathbf{A}_j$, then

$$I(S; \mathbf{A}_j) = I_{\min}(S; \mathbf{A}_j) = I_{\min}(S; \mathbf{A}_i, \mathbf{A}_j).$$

We also know that

$$I(S; \mathbf{A}_i) = I_{\min}(S; \mathbf{A}_i) \geq I_{\min}(S; \mathbf{A}_i, \mathbf{A}_j),$$

so we conclude that $I(S; \mathbf{A}_i) \geq I(S; \mathbf{A}_j)$. □

The following corollary states that complete dependence (between sources) implies complete redundancy.

**Corollary 4.4.** *If* $\mathbf{A}_j = f(\mathbf{A}_i)$*, then*

$$I_{\min}(S; \mathbf{A}_i, \mathbf{A}_j) = I(S; \mathbf{A}_j).$$

*Proof.* If $\mathbf{A}_j = f(\mathbf{A}_i)$, then $S \to \mathbf{A}_i \to \mathbf{A}_j$. □

However, it is worth noting that, aside from the connection in Corollary 4.4, dependence between sources and redundancy are in fact completely dissociable. That is, neither does dependence imply redundancy, nor the converse, as the following simple examples illustrate. Let $R_1$ and $R_2$ be binary and uniformly distributed. If $R_1 = R_2$ and $S$ is a degenerate random variable, then $I(R_1; R_2) > 0$ and $I_{\min}(S; R_1, R_2) = 0$, so that dependence does not imply redundancy. On the other hand, if $R_1$ and $R_2$ are independent and $S = R_1 \wedge R_2$, then $I_{\min}(S; R_1, R_2) > 0$ and $I(R_1; R_2) = 0$, so that redundancy does not imply dependence. Considering instead conditional dependence, if $R_1$ and $R_2$ are independent and $S = R_1 \oplus R_2$, then $I(R_1; R_2|S) > 0$ and $I_{\min}(S; R_1, R_2) = 0$, so that conditional dependence does not imply redundancy. Conversely, if $S = R_1 = R_2$, then $I_{\min}(S; R_1, R_2) > 0$ and $I(R_1; R_2|S) = 0$, so that redundancy does not imply conditional dependence.

## 4.4 Redundancy Lattice

What are the distinct ways in which collections of sources might contribute redundant information? Formally, answering this question means identifying the domain of $I_\cap$. Thus far, we have implicitly assumed that the natural domain is the collection of all possible sets of sources, but in fact this domain can be greatly simplified. Given sources $\mathbf{A}_i$ and $\mathbf{A}_j$ with $\mathbf{A}_i \subseteq \mathbf{A}_j$, Axiom 4.3 states that the redundancy for $\mathbf{A}_i$ and $\mathbf{A}_j$ is equivalent to the redundancy when $\mathbf{A}_j$ is removed:

$$I_\cap(S; \mathbf{A}_i, \mathbf{A}_j, \ldots) = I_\cap(S; \mathbf{A}_i, \ldots).$$

Applying this property recursively, it follows that for any collection of sources where some are supersets of others, the redundancy for that collection is equivalent to the redundancy when all supersets are removed. Thus, the domain for $I_\cap$ can be reduced to the collection of all sets of sources such that no source is a superset of any other:

$$\mathcal{A}(\mathbf{R}) = \{\alpha \in \mathcal{P}_1(\mathcal{P}_1(\mathbf{R})) : \forall \mathbf{A}_i, \mathbf{A}_j \in \alpha, \mathbf{A}_i \not\subset \mathbf{A}_j\}, \tag{4.3}$$

where $\mathcal{P}_1(\mathbf{R}) = \mathcal{P}(\mathbf{R}) \setminus \{\emptyset\}$ is the set of all nonempty subsets of $\mathbf{R}$. Formally, $\mathcal{A}(\mathbf{R})$ corresponds to the set of all *antichains* on the inclusion lattice $\langle \mathcal{P}(\mathbf{R}), \subseteq \rangle$, excluding the empty set[1]. The cardinality of this set is given by the $n$-th Dedekind number, which for $n = 1, 2, 3, \ldots$ is $1, 4, 18, 166, 7579, \ldots$ ( [34], p. 273). Henceforth, we will denote elements of $\mathcal{A}(\mathbf{R})$, corresponding to collections of sources, with bracketed expressions containing only the indices for each source. For instance, $\{\{R_1, R_2\}\}$ will be $\{12\}$, $\{\{R_1\}, \{R_2, R_3\}\}$ will be $\{1\}\{23\}$, and so forth.

The possibilities for redundancy are also naturally structured, which can be shown by

[1] Basic concepts from lattice theory are reviewed in Appendix A.

extending the same line of reasoning to define a partial ordering on the elements of $\mathcal{A}(\mathbf{R})$. Consider two collections of sources, $\alpha, \beta \in \mathcal{A}(\mathbf{R})$, where for each source $\mathbf{B} \in \beta$ there exists a source $\mathbf{A} \in \alpha$ such that $\mathbf{A} \subseteq \mathbf{B}$. This means that for each $\mathbf{B} \in \beta$ there is an $\mathbf{A} \in \alpha$ such that $\mathbf{A}$ provides no more information than $\mathbf{B}$. The redundant information shared by all $\mathbf{B} \in \beta$ must therefore at least include any redundant information shared by all $\mathbf{A} \in \alpha$. Thus, we can define a partial order $\preccurlyeq$ over the elements of $\mathcal{A}(\mathbf{R})$ such that one element (collection of sources) precedes another if, and only if, the latter provides any redundant information that the former provides:

$$\forall \alpha, \beta \in \mathcal{A}(\mathbf{R}), (\alpha \preccurlyeq \beta \Leftrightarrow \forall \mathbf{B} \in \beta, \exists \mathbf{A} \in \alpha, \mathbf{A} \subseteq \mathbf{B}). \tag{4.4}$$

Applying this ordering to the elements of $\mathcal{A}(\mathbf{R})$ produces a *redundancy lattice*, in which a higher element provides at least as much redundant information as a lower one (Figure 4.3). The fact that $\langle \mathcal{A}(\mathbf{R}), \preccurlyeq \rangle$ forms a lattice is proven in [37], where the corresponding lattice is denoted $\langle \mathcal{A}(X), \preccurlyeq' \rangle$ (see also [38]). As shown in [37], the meet ($\wedge$) and join ($\vee$) for this lattice are given by

$$\alpha \wedge \beta = \underline{\alpha \cup \beta} \tag{4.5}$$

and

$$\alpha \vee \beta = \underline{\uparrow \alpha \cap \uparrow \beta}. \tag{4.6}$$

From the redundancy lattice, it is possible to read off some of the properties of $I_\cap$ noted earlier. For instance, the property that redundancy is maximized by self-redundancy corresponds to the fact that any node representing a single source is higher in the lattice than any other node involving that source. For example, in Figure 4.3(b), the node labeled $\{12\}$, corresponding to self-redundancy for the source $\{R_1, R_2\}$, is higher than the nodes labeled $\{12\}\{13\}$, $\{12\}\{13\}\{23\}$, and $\{3\}\{12\}$. Another useful property of $I_\cap$ relates to the top and

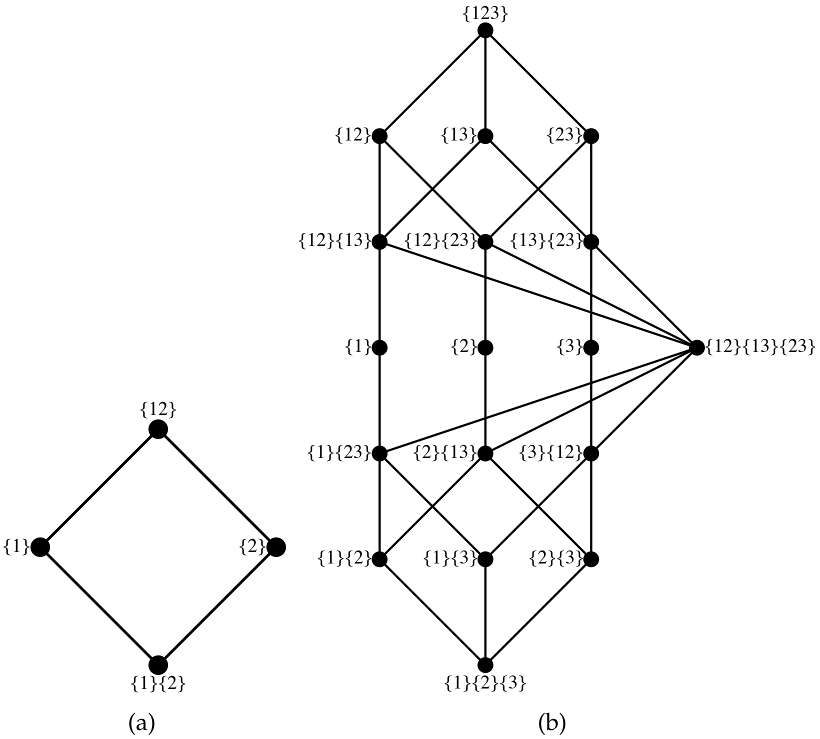


Figure 4.3: Redundancy lattice for (a) 3 and (b) 4 variables.

bottom nodes of the lattice. The top node corresponds to the self-redundancy for **R**, reflecting the fact that $I_\cap$ is bounded from above by the total information from **R**. At the other end of the spectrum, the bottom element corresponds to the redundant information that each individual variable in **R** provides, with all other possibilities for redundancy falling between these two extremes.

## 4.5 Partial Information Decomposition

The redundant information associated with each node of the redundancy lattice includes, but is not limited to, the redundant information provided by all nodes lower in the lattice. Thus, moving from node to node up the lattice, $I_\cap$ can be thought of as a kind of "cumulative information function," effectively integrating the information provided by increasingly inclusive collections of sources. Next, we define an inverse of $I_\cap$ called the *partial information (PI) function*. The relationship between $I_\cap$ and the PI-function is analogous to that between the integral and the derivative from elementary calculus: moving up the redundancy lattice, $I_\cap$ integrates the information associated with individual nodes, while the PI-function measures the information contributed by each individual node. Thus, whereas $I_\cap$ quantifies cumulative information, the PI-function measures the partial information contributed uniquely by each particular collection of sources. This partial information will form the atoms into which we decompose the total information that $\mathbf{R}$ provides about $S$.

For a collection of sources $\alpha \in \mathcal{A}(\mathbf{R})$, the PI-function, denoted $I_\partial$, is defined implicitly by

$$I_\cap(S;\alpha) = \sum_{\beta \preccurlyeq \alpha} I_\partial(S;\beta). \tag{4.7}$$

This equation is an instance of the Möbius inversion formula [192, 219], a basic concept from algebraic combinatorics that is analogous to the fundamental theorem of calculus. From this relationship, it is clear that $I_\partial$ can be calculated recursively as

$$I_\partial(S;\alpha) = I_\cap(S;\alpha) - \sum_{\beta \prec \alpha} I_\partial(S;\beta). \tag{4.8}$$

Thus, in words, $I_\partial(S;\alpha)$ quantifies the information provided redundantly by the sources of $\alpha$ that is not provided by any simpler collection of sources (i.e., any $\beta$ lower than $\alpha$ on the redundancy lattice). The following theorem provides a closed-form expression for $I_\partial$.

**Theorem 4.4.**

$$I_\partial(S;\alpha) = I_\cap(S;\alpha) - \sum_{k=1}^{|\alpha^-|}(-1)^{k-1} \sum_{\substack{\mathcal{B}\subseteq\alpha^- \\ |\mathcal{B}|=k}} I_\cap(S; \bigwedge \mathcal{B}).$$

*Proof.* For $\mathcal{B} \subseteq \mathcal{A}(\mathbf{R})$, define the set-additive function $f$ as

$$f(\mathcal{B}) = \sum_{\beta \in \mathcal{B}} I_\partial(S; \beta).$$

From Equation (4.7), it follows that $I_\cap(S; \alpha) = f(\downarrow \alpha)$ and

$$\begin{aligned} I_\partial(S; \alpha) &= f(\downarrow \alpha) - f(\dot{\downarrow} \alpha) \\ &= f(\downarrow \alpha) - f(\bigcup_{\beta \in \alpha^-} \downarrow \beta). \end{aligned}$$

Applying the inclusion-exclusion principle ( [219], p. 64), we have

$$= f(\downarrow \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}| = k}} f(\bigcap_{\gamma \in \mathcal{B}} \downarrow \gamma)$$

and it is a basic result of lattice theory that for any lattice $L$ and $A \subseteq L$, $\bigcap_{a \in A} \downarrow a = \downarrow (\bigwedge A)$ ( [44], p. 57), so we have

$$\begin{aligned} &= f(\downarrow \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}| = k}} f(\downarrow (\bigwedge \mathcal{B})) \\ &= I_\cap(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}| = k}} I_\cap(S; \bigwedge \mathcal{B}). \end{aligned}$$

□

The expression for $I_\partial$ established in Theorem 4.4 is best understood in terms of the inclusion-exclusion principle [219], which serves as the basis for its derivation. To illustrate this idea, the calculation of $I_\partial(S; \{12\})$ for $\mathbf{R} = \{R_1, R_2\}$ is depicted in Figure 4.4. First, $I_\cap(S; \{12\})$ is included, which equals the sum of $I_\partial(S; \alpha)$ for all $\alpha \in \downarrow \{12\}$ (indicated by

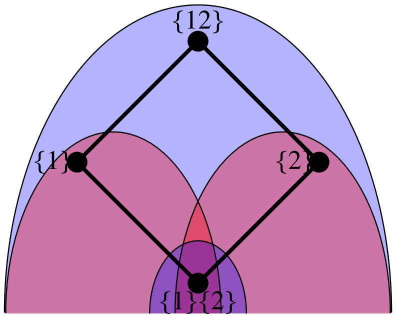


Figure 4.4: Inclusion-exclusion formula for $I_\partial$. The calculation of $I_\partial(S; \{12\})$ for $\mathbf{R} = \{R_1, R_2\}$ is depicted. Blue and red regions correspond to $I_\cap$ terms that are added and subtracted, respectively.

the large blue region). Next, $I_\cap(S; \beta)$ is subtracted off for each $\beta \in \{12\}^-$ (represented by the two red regions). However, this step subtracts twice the $I_\partial$ term that is included in the down-sets of both elements in $\{12\}^-$ (the term $\{1\}\{2\}$), so the next step of the summation adds this terms back in (the small blue region). In general, this pattern then continues by alternately adding and subtracting $I_\cap$ terms in order to capture the $I_\partial$ terms that are common to the down-sets of increasingly large subsets of $\alpha^-$. Hopefully it is clear how this calculation relates to the standard use of inclusion-exclusion to calculate the cardinality of the union of several sets by adding up their individual cardinalities, subtracting their pairwise intersections, adding in their three-way intersections, and so forth.

The PI-function is nonnegative if, and only if, $I_\cap$ is *totally monotone*[2] [81, 121], meaning that

$$I_\cap(S; \bigvee_{1 \le j \le k} \alpha_j) \ge \sum_{\emptyset \ne J \subseteq \{1,\ldots,k\}} (-1)^{|J|-1} I_\cap(S; \bigwedge_{j \in J} \alpha_j) \tag{4.9}$$

for $k \ge 2$ and every $k$-element subset of $\mathcal{A}(\mathbf{R})$. Total monotonicity is a stronger form of supermodularity[3], which, along with the dual concept of submodularity, has recently been

[2] The result that the Möbius inverse of a lattice function is nonnegative if, and only if, that function is totally monotone is proven in [74] and discussed further in [191]. Totally monotone functions are also called totally positive [74, 191] or strongly increasing [142] functions.

[3] Supermodularity means that Equation (4.9) holds for $k = 2$.

discussed as a central or even defining property of information [147,152,220,221]. For example, in [221] submodularity is incorporated into an axiomatic definition of entropy, and it is shown that (conditional) mutual information is nonnegative if, and only if, entropy is submodular (recall Equation (2.8) and the discussion thereof). Thus, there is an attractive parallel between the relationships (nonnegative mutual information ↔ submodular entropy) and (nonnegative partial information ↔ totally monotone redundancy). Later, we will show that $I_{\min}$ has the desired property of total monotonicity, so that its corresponding PI-function decomposes $I(S;\mathbf{R})$ into a sum of nonnegative PI-terms.

The decomposition of $I(S;\mathbf{A})$ into a sum of PI-terms follows from

$$I(S;\mathbf{A}) = I_{\cap}(S;\mathbf{A}) = \sum_{\beta \preccurlyeq \{\mathbf{A}\}} I_{\partial}(S;\beta). \tag{4.10}$$

Thus, in terms of the redundancy lattice, the total information from $\mathbf{A}$ decomposes into the sum of all PI-terms for elements in $\downarrow \{\mathbf{A}\}$. The chain rule for information (Equation (2.19)) yields a similar decomposition for the conditional mutual information:

$$I(S;\mathbf{A}|\mathbf{B}) = \sum_{\substack{\beta \preccurlyeq \{\mathbf{A}\cup\mathbf{B}\} \\ \beta \not\preccurlyeq \{\mathbf{B}\}}} I_{\partial}(S;\beta). \tag{4.11}$$

Thus, the conditional mutual information decomposes into the sum of all PI-terms for elements in $\downarrow \{\mathbf{A} \cup \mathbf{B}\}$ that are not in $\downarrow \{\mathbf{B}\}$. Note that, since we are dealing with partial orders, it is necessary to use two inequalities, $\beta \preccurlyeq \{\mathbf{A} \cup \mathbf{B}\}$ and $\beta \not\preccurlyeq \{\mathbf{B}\}$, instead of a combined inequality, $\{\mathbf{B}\} \prec \beta \preccurlyeq \{\mathbf{A} \cup \mathbf{B}\}$.

For the 3-variable case, $\mathbf{R} = \{R_1, R_2\}$, Equation (4.10) yields

$$I(S;R_1) = I_{\partial}(S;\{1\}) + I_{\partial}(S;\{1\}\{2\}) \tag{4.12}$$

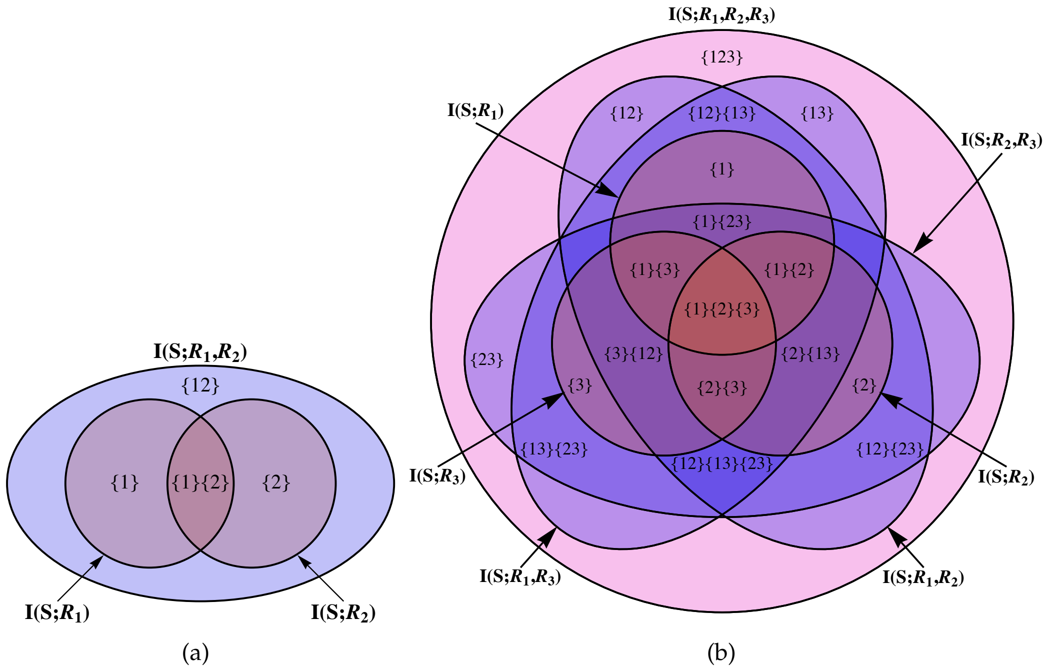


Figure 4.5: Partial information diagrams for (a) 3 and (b) 4 variables.

and

$$I(S;R_1,R_2) = I_\partial(S;\{1\}) + I_\partial(S;\{2\}) + I_\partial(S;\{1\}\{2\}) + I_\partial(S;\{12\}). \qquad (4.13)$$

The relationship between these equations can be represented as a *PI-diagram* (Figure 4.5(a)), which illustrates the set-theoretic breakdown of total information into PI-terms. Also, comparing this diagram with Figure 4.2 makes immediately clear the meaning of each PI-term. First, from Theorem 4.4, we have that $I_\partial(S;\{1\}\{2\}) = I_\cap(S;\{1\}\{2\})$, corresponding to the redundancy for $R_1$ and $R_2$. The unique information for $R_1$ is given by $I_\partial(S;\{1\}) = I(S;R_1) - I_\cap(S;\{1\}\{2\})$, which is the total information from $R_1$ minus the redundancy with $R_2$, and likewise for $R_2$. Finally, the additional information provided by the combination of $R_1$ and $R_2$ is given by $I_\partial(S;\{12\})$, corresponding to their synergy.

The general structure of PI-diagrams becomes clear when we consider the decomposition for four variables (Figure 4.5(b)). First, note that all of the possibilities for three variables are again present for four. In particular, each element of $\mathbf{R}$ can provide unique information (regions labeled {1}, {2}, and {3}), information redundantly with one other variable ({1}{2}, {1}{3}, and {2}{3}), or information synergistically with one other variable ({12}, {13}, and {23}). Additionally, information can be provided redundantly by all three variables ({1}{2}{3}) or provided by their three-way synergy ({123}). More interesting are the new kinds of terms representing combinations of redundancy and synergy. For instance, the regions marked {1}{23}, {2}{13}, and {3}{12} represent information that is available redundantly from either one variable considered individually or the other two considered together. Or, for instance, the region labeled {12}{13}{23} represents the information provided redundantly by the three possible two-way synergies. In general, the PI-atom for a collection of sources corresponds to the information provided redundantly by the synergies of all sources in the collection. This point also clarifies our earlier claim that unique information is best thought of as a degenerate case: unique information corresponds to the combination of first-order redundancy and first-order synergy.

In general, the PI-diagram for $S$ and $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$ consists of the following (Figure 4.6). First, for each element $R_i \in \mathbf{R}$ there is a region corresponding to $I(S; R_i)$. Then, for every subset $\mathbf{A}$ of $\mathbf{R}$ with two or more elements, $I(S; \mathbf{A})$ is depicted as a region containing $I(S; A)$ for all $A \in \mathbf{A}$ but not coextensive with $\bigcup_{A \in \mathbf{A}} I(S; A)$. The difference between $I(S; \mathbf{A})$ and $\bigcup_{A \in \mathbf{A}} I(S; A)$ represents the synergy for $\mathbf{A}$, the information gained from the combined knowledge of all elements in $\mathbf{A}$ that is not available from any subset. In addition, regions of the diagram intersect generically, representing all possibilities for redundancy. In total, a PI-diagram is composed of the $n$-th Dedekind number [34] of PI-atoms, same as the cardinality of $\mathcal{A}(\mathbf{R})$. As described above, each PI-atom represents the redundancy of synergies for a particular collection of sources, corresponding to one distinct way for the components of $\mathbf{R}$ to contribute information about $S$.

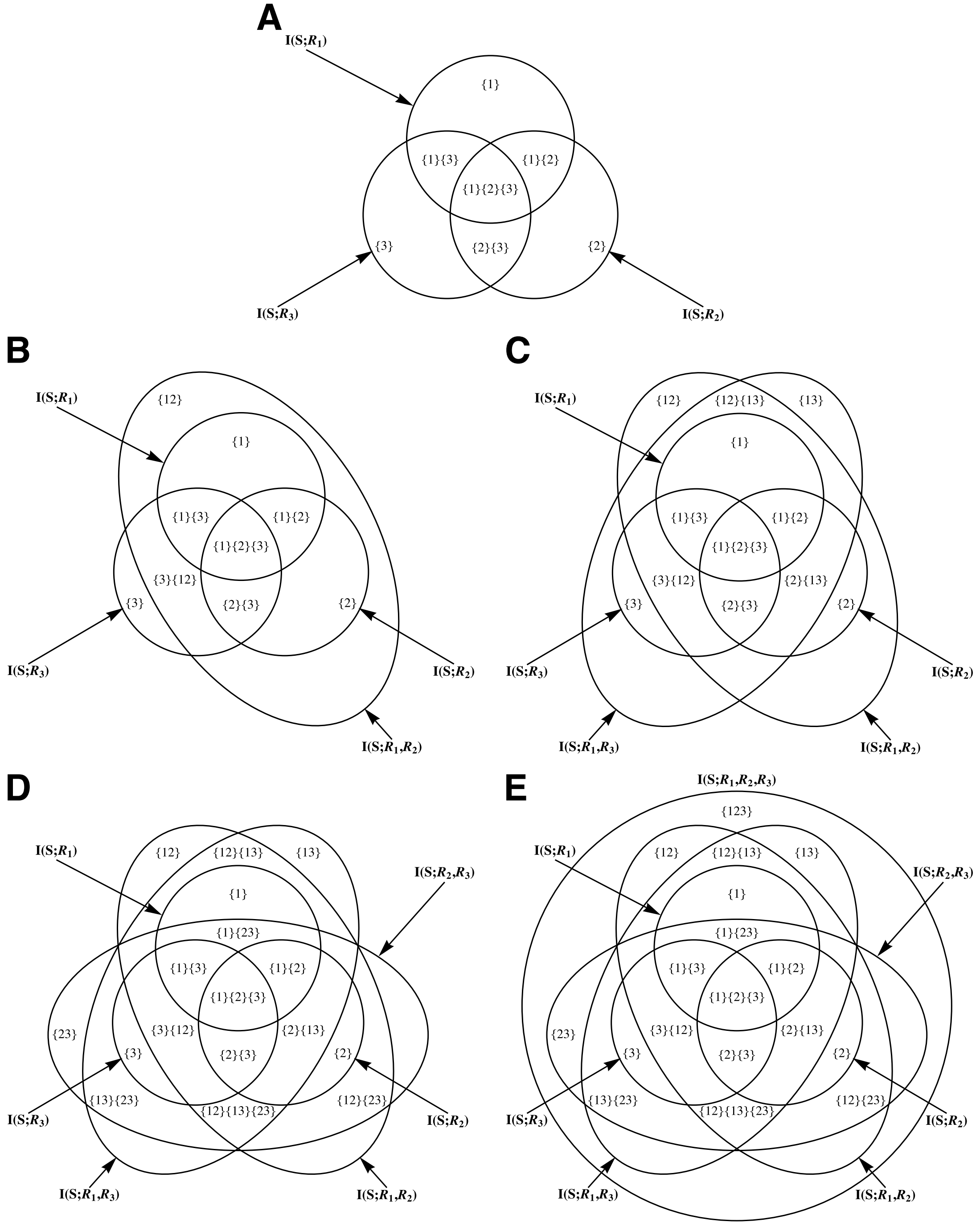


Figure 4.6: Constructing a PI-diagram. (A) For each element $R_i \in \mathbf{R}$ there is a region corresponding to $I(S; R_i)$. (B-E) For each subset $\mathbf{A}$ of $\mathbf{R}$ with two or more elements, $I(S; \mathbf{A})$ is depicted as a region containing $I(S; A)$ for all $A \in \mathbf{A}$ but not coextensive with $\bigcup_{A \in \mathbf{A}} I(S; A)$.

PI-diagrams represent geometrically the underlying set-theoretic structure of multivariate information. In this context, $I_\cap$ is formally analogous to set intersection, with $I_\cap(S; \mathbf{A}_1, \ldots, \mathbf{A}_k)$ corresponding to the region $\bigcap_i I(S; \mathbf{A}_i)$. This relationship also connects the redundancy lattice and PI-diagram for $n$ variables: for $\alpha, \beta \in \mathcal{A}(\mathbf{R})$, $\alpha$ is lower than $\beta$ in the redundancy lattice if, and only if, $\bigcap_{\mathbf{A}\in\alpha} I(S; \mathbf{A})$ is a subset of $\bigcap_{\mathbf{B}\in\beta} I(S; \mathbf{B})$ in the PI-diagram. If $I_\cap$ has a nonnegative PI-function (as is the case, for instance, with $I_{\text{min}}$), then PI-diagrams represent a measure, in the formal sense of measure theory, over sets representing the information from each source.

It is important to contrast the measure represented by PI-diagrams with that represented by *I-diagrams*, which were described in Section 2.3 (see, in particular, Figure 2.1). The crucial difference between PI-diagrams and I-diagrams is that I-diagrams represent *signed* measures, meaning that the area of regions in an I-diagram can represent negative values. This is equivalent to the property that interaction information is sometimes negative, an idea that we revisit in the next section.

Now we turn from the general PI-function that is the inverse of $I_\cap$ to the specific PI-function that is the inverse of $I_{\text{min}}$. We denote the PI-function corresponding to $I_{\text{min}}$ by $\Pi$, so that

$$I_{\text{min}}(S; \alpha) = \sum_{\beta \preccurlyeq \alpha} \Pi(S; \beta). \tag{4.14}$$

**Theorem 4.5.** $\Pi$ *is nonnegative.*

*Proof.* If $\alpha = \bot$, $\Pi(S; \alpha) = I_{\text{min}}(S; \alpha)$ and $\Pi(S; \alpha) \geq 0$ follows from the nonnegativity of $I_{\text{min}}$. To prove it for $\alpha \neq \bot$, we proceed by contradiction. Assume there exists $\alpha \in \mathcal{A}(\mathbf{R}) \setminus \{\bot\}$ such that $\Pi(S; \alpha) < 0$. Beginning with the simpler expression for $\Pi(S; \alpha)$ proven in Theorem 4.6 and combining summations yields

$$\Pi(S; \alpha) = \sum_s p(s) \{\min_{\mathbf{A}\in\alpha} I(S = s; \mathbf{A}) - \max_{\beta\in\alpha^-} \min_{\mathbf{B}\in\beta} I(S = s; \mathbf{B})\}.$$

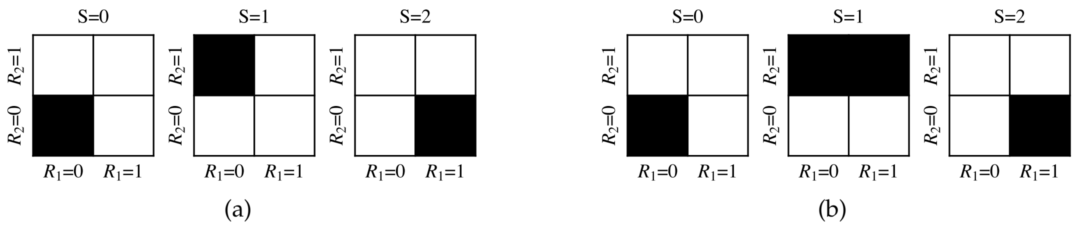


Figure 4.7: Examples of mixed synergy and redundancy. Each set of tiled squares represents a probability distribution for $S \in \{0, 1, 2\}$ and $R_1, R_2 \in \{0, 1\}$. Black tiles represent equiprobable outcomes and white tiles represent zero-probability outcomes.

From this equation, it is clear that there must exist $\beta \in \alpha^-$ such that for all $\mathbf{B} \in \beta$, $I(S = s; \mathbf{A}) < I(S = s; \mathbf{B})$ for some outcome $s \in S$ and some $\mathbf{A} \in \alpha$. Thus, from Lemma 2.2, there does not exist $\mathbf{B} \in \beta$ such that $\mathbf{B} \subseteq \mathbf{A}$. However, since $\beta \prec \alpha$ by definition, there exists $\mathbf{B} \in \beta$ such that $\mathbf{B} \subseteq \mathbf{A}$. □

As a first example of PI-decomposition using the PI-function Π, consider the distribution in Figure 4.7(a). From the symmetry of the distribution, it is clear that $R_1$ and $R_2$ must provide the same amount of information about $S$. Indeed, this is easily verified, with $I(S; R_1) = I(S; R_2) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}$. However, it is also clear that $R_1$ and $R_2$ provide information about different outcomes of $S$. In particular, given knowledge of $R_1$, one can determine conclusively whether or not outcome $S = 2$ occurs (which is not the case for $R_2$), and likewise for $R_2$ and outcome $S = 1$. This is captured by the fact that $\Pi(S; \{1\}) = \Pi(S; \{2\}) = \frac{1}{3}$, meaning that $R_1$ and $R_2$ each provide $\frac{1}{3}$ bits of unique information about $S$. The redundant information, $\Pi(S; \{1\}\{2\}) = \log 3 - \log 2$, reflects the fact that both $R_1$ or $R_2$ reduce uncertainty about $S$ from three equally likely outcomes to two. Finally, $R_1$ and $R_2$ also provide $\frac{1}{3}$ bits of synergistic information, $\Pi(S; \{12\}) = \frac{1}{3}$. This value corresponds to the fact that $R_1$ and $R_2$ together uniquely determine whether or not outcome $S = 0$ occurs, which is not true for $R_1$ or $R_2$ alone.

Note that, unlike mutual information or interaction information, partial information is *not* symmetric. For instance, the synergistic information that $R_1$ and $R_2$ provide about $S$

is not in general equal to the synergistic information that $S$ and $R_2$ provide about $R_1$. This property is also illustrated by the example in Figure 4.7(a). Given knowledge of $S$, one can uniquely determine the outcome of $R_1$ (and $R_2$), so that $S$ provides complete information about both variables. Thus, it is not possible for the combination of $S$ and $R_2$ to provide any additional synergistic information about $R_1$, since there is no remaining uncertainty about $R_1$ when $S$ is known. In contrast, as was just noted, $R_1$ and $R_2$ provide $\frac{1}{3}$ bits of synergistic information about $S$. This asymmetry accounts for our decision to focus on information *about* a particular variable $S$ throughout, since in general the analysis will differ depending on the variable of interest. Note that total information is also asymmetric in this sense[4], i.e., in general $I(S;R_1,R_2) \neq I(R_1;S,R_2)$.

Finally, we note two additional properties that are useful for the practical computation of PI-terms. Both properties make use of the following relationship between $\max$ and $\min$.

**Lemma 4.1** (Maximum-minimums identity). *For any set of numbers $A$,*

$$\max A = \sum_{k=1}^{|A|} (-1)^{k-1} \sum_{\substack{B \subseteq A \\ |B|=k}} \min B,$$

*or conversely,*

$$\min A = \sum_{k=1}^{|A|} (-1)^{k-1} \sum_{\substack{B \subseteq A \\ |B|=k}} \max B.$$

*Proof.* It is proven in a number of introductory texts, e.g. [190]. □

First, the following theorem provides a simpler expression for $\Pi$.

**Theorem 4.6.**

$$\Pi(S;\alpha) = I_{\min}(S;\alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S=s;\mathbf{B}).$$

[4] Though, of course, it is symmetric in the sense that $I(S;R_1,R_2) = I(R_1,R_2;S)$.

*Proof.*

Rewriting Theorem 4.4, $\Pi(S;\alpha)$ can be written as

$$
\begin{aligned}
& I_{\min}(S;\alpha) - \sum_{k=1}^{|\alpha^-|}(-1)^{k-1} \sum_{\substack{\mathcal{B}\subseteq\alpha^- \\ |\mathcal{B}|=k}} \sum_{s} p(s) \min_{\mathbf{B}\in\bigwedge\mathcal{B}} I(S=s;\mathbf{B}) \\
&= I_{\min}(S;\alpha) - \sum_{s} p(s) \sum_{k=1}^{|\alpha^-|}(-1)^{k-1} \sum_{\substack{\mathcal{B}\subseteq\alpha^- \\ |\mathcal{B}|=k}} \min_{\mathbf{B}\in\bigwedge\mathcal{B}} I(S=s;\mathbf{B})
\end{aligned}
$$

and by Lemma 2.2 and Equation (4.5),

$$
= I_{\min}(S;\alpha) - \sum_{s} p(s) \sum_{k=1}^{|\alpha^-|}(-1)^{k-1} \sum_{\substack{\mathcal{B}\subseteq\alpha^- \\ |\mathcal{B}|=k}} \min_{\beta\in\mathcal{B}} \min_{\mathbf{B}\in\beta} I(S=s;\mathbf{B}).
$$

Then, applying Lemma 4.1 we have

$$
= I_{\min}(S;\alpha) - \sum_{s} p(s) \max_{\beta\in\alpha^-} \min_{\mathbf{B}\in\beta} I(S=s;\mathbf{B}).
$$

□

Unlike Theorem 4.4, this expression for $\Pi(S;\alpha)$ only requires knowledge of the nodes *immediately* below $\alpha$ in the redundancy lattice (the set $\alpha^-$). Second, in the same way that $I_{\min}$ acts like set intersection for PI-diagrams, $I_{\max}$ acts like set union, where $I_{\max}$ is defined exactly the same as $I_{\min}$ except substituting $\max$ for $\min$. This is captured by the following theorem, which links $I_{\max}$ and $I_{\min}$ by an inclusion-exclusion relationship that is analogous to that between $\cup$ and $\cap$.

**Theorem 4.7.**

$$I_{\max}(S;\alpha) = \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta| = k}} I_{\min}(S;\beta).$$

*Proof.* Using Lemma 4.1, $I_{\max}$ can be written as

$$I_{\max}(S;\alpha) = \sum_{s} p(s) \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta| = k}} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}),$$

and rearranging the order of summations,

$$\begin{aligned} &= \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta| = k}} \sum_{s} p(s) \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \\ &= \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta| = k}} I_{\min}(S;\beta). \end{aligned}$$

□

Thus, $I_{\max}(S; \mathbf{A}_1, \ldots, \mathbf{A}_k)$ corresponds to the PI-diagram region $\bigcup_i I(S; \mathbf{A}_i)$, which has the appealing interpretation that the informational union of several sources equals the maximum information that is available from any one source. This correspondence between $I_{\min}/I_{\max}$ and $\cap/\cup$ means that expressions for PI-regions can be derived using the rich operations of set theory. For example, the synergy for all $n$ variables in $\mathbf{R}$ is conveniently given by

$$I(S;\mathbf{R}) - I_{\max}(S; \{\mathbf{A} \subset \mathbf{R} : |\mathbf{A}| = n-1\}). \tag{4.15}$$

Likewise, the unique information from a single $R_i$ is given simply by

$$I(S;R_i) - I_{\min}(S; R_i, \mathbf{R} \setminus R_i). \tag{4.16}$$

## 4.6 Why Interaction Information is Sometimes Negative

We next show how PI-decomposition can be used to understand the conditions under which interaction information, the standard generalization of mutual information to multivariate interactions, is negative. The interaction information [153] for three variables is given by

$$I(S;R_1;R_2) = I(S;R_1|R_2) - I(S;R_1) \tag{4.17}$$

and for $n > 3$ variables is defined recursively as

$$\begin{aligned} I(S;R_1;R_2;\ldots;R_{n-1}) =& I(S;R_1;R_2;\ldots;R_{n-2}|R_{n-1}) \\ & - I(S;R_1;R_2;\ldots;R_{n-2}) \end{aligned} \tag{4.18}$$

where the conditional interaction information is defined by simply including the conditioning in all terms of the original definition. Interaction information is symmetric for all permutations of its arguments, and is traditionally interpreted as the information shared by all $n$ variables beyond that which is shared by any subset of those variables [21,92,226, 236,259,262].

For 3-variable interaction information, a positive value is naturally interpreted as indicating a situation in which any one variable of the system enhances the correlation between the other two. For example, a positive value for Equation (4.17) means that knowledge of $R_2$ enhances the correlation between $S$ and $R_1$ (and likewise for all other variable permutations). Thus, in the terminology used here, a positive value for $I(S;R_1;R_2)$ signals the presence of synergy. On the other hand, a negative value for $I(S;R_1;R_2)$ indicates a situation in which any one variable accounts for or “explains away” [172] the correlation between the other two. In other words, a negative value for $I(S;R_1;R_2)$ indicates redundancy. Indeed, $I(S;R_1;R_2)$ is a widely used measure of synergy and redundancy in

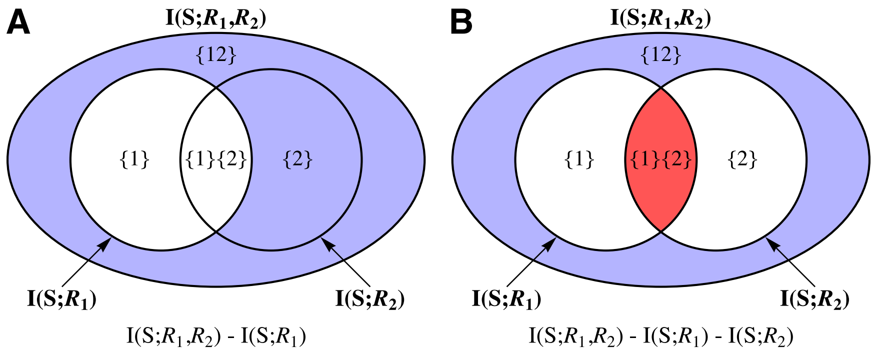


Figure 4.8: PI-decomposition of $I(S;R_1;R_2)$. (A-B) Term-by-term calculation of $I(S;R_1;R_2) = I(S;R_1,R_2) - I(S;R_1) - I(S;R_2)$. Blue and red regions represent PI-terms that are added and subtracted, respectively.

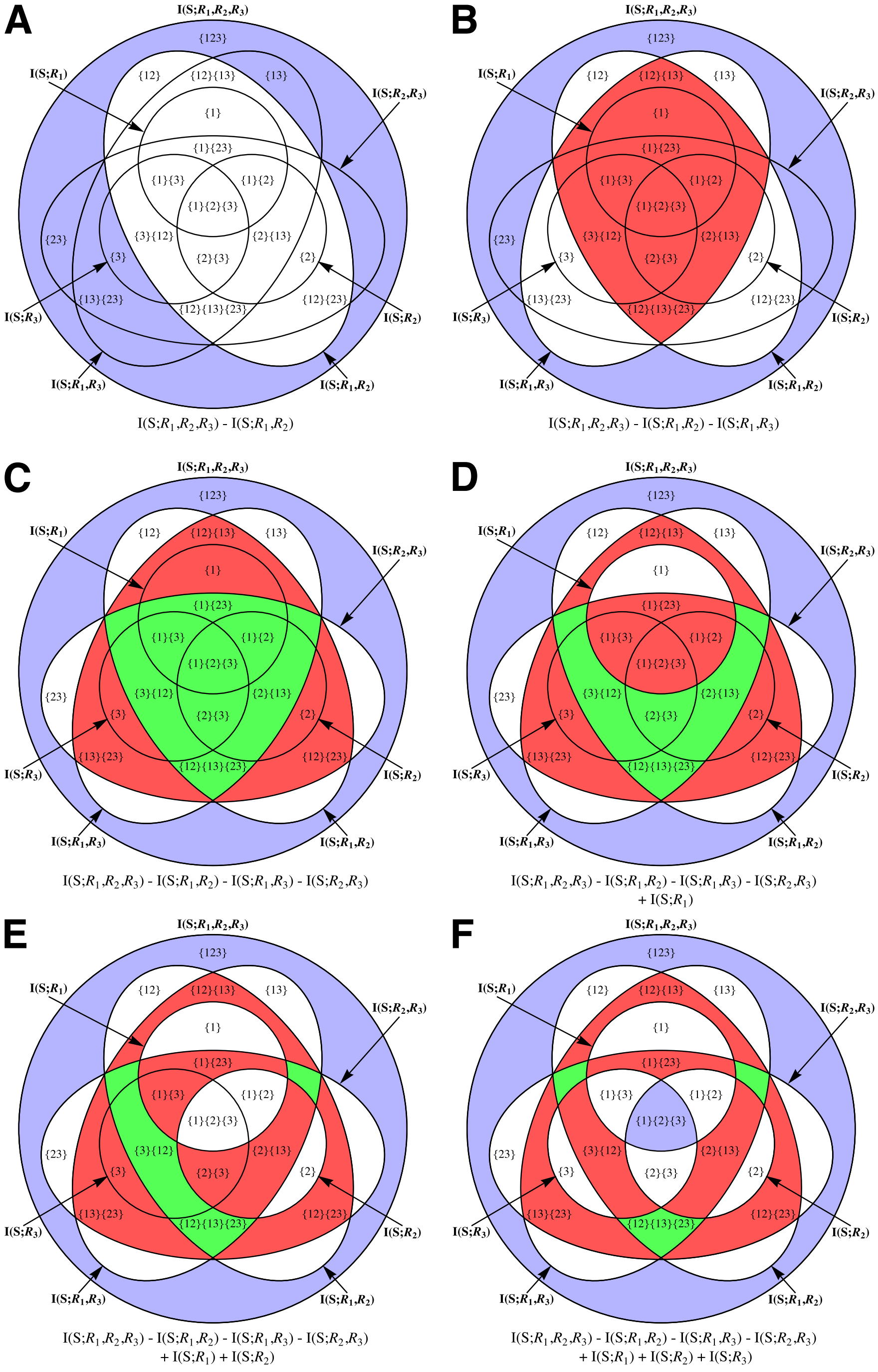
neuroscience, where it is interpreted in exactly this way [25,131,169,196].

The PI-decomposition for 3-variable interaction information (Figure 4.8) confirms this interpretation. From Equations (4.12), (4.13) and (4.17), we have that $I(S;R_1;R_2)$ is equal to the difference between the synergistic and the redundant information, i.e.,

$$I(S;R_1;R_2) = \Pi(S;\{12\}) - \Pi(S;\{1\}\{2\}). \tag{4.19}$$

Thus, it is indeed the case that positive values indicate synergy and negative values indicate redundancy.

However, the PI-decomposition also makes clear that $I(S;R_1;R_2)$ confounds redundancy and synergy, with the meaning of interaction information ambiguous for any system that exhibits a mixture of the two (cf. [108], who suggest the possibility of mixed redundancy and synergy, but without attempting to disentangle them). For instance, consider again the example in Figure 4.7(a). As described earlier, $R_1$ and $R_2$ provide $\log 3 - \log 2$ bits of redundant information and $\frac{1}{3}$ bits of synergistic information. Thus, $I(S;R_1;R_2)$ is negative because there is more redundancy than synergy, despite the fact that the system clearly exhibits synergistic interactions. As a second example, consider the distribution in



I(S;$R_1$,$R_2$,$R_3$) - I(S;$R_1$,$R_2$)

I(S;$R_1$,$R_2$,$R_3$) - I(S;$R_1$,$R_2$) - I(S;$R_1$,$R_3$)

I(S;$R_1$,$R_2$,$R_3$) - I(S;$R_1$,$R_2$) - I(S;$R_1$,$R_3$) - I(S;$R_2$,$R_3$)

I(S;$R_1$,$R_2$,$R_3$) - I(S;$R_1$,$R_2$) - I(S;$R_1$,$R_3$) - I(S;$R_2$,$R_3$)
+ I(S;$R_1$)

I(S;$R_1$,$R_2$,$R_3$) - I(S;$R_1$,$R_2$) - I(S;$R_1$,$R_3$) - I(S;$R_2$,$R_3$)
+ I(S;$R_1$) + I(S;$R_2$)

I(S;$R_1$,$R_2$,$R_3$) - I(S;$R_1$,$R_2$) - I(S;$R_1$,$R_3$) - I(S;$R_2$,$R_3$)
+ I(S;$R_1$) + I(S;$R_2$) + I(S;$R_3$)

Figure 4.9: PI-decomposition of $I(S;R_1;R_2;R_3)$. (A-F) Term-by-term calculation of $I(S;R_1;R_2;R_3) = I(S;R_1,R_2,R_3) - I(S;R_1,R_2) - I(S;R_1,R_3) - I(S;R_2,R_3) + I(S;R_1) + I(S;R_2) + I(S;R_3)$. Blue and red regions represent PI-terms that are added and subtracted, respectively. Green regions represent PI-terms that are subtracted twice.

Figure 4.7(b). In this case, $R_1$ and $R_2$ provide $\frac{1}{2}$ bits of redundant information, since $R_1$ or $R_2$ both reduce uncertainty about the outcomes $S = 0$ and $S = 2$. $R_1$ and $R_2$ also provide $\frac{1}{2}$ bits of synergistic information, reflecting the fact that $R_1$ and $R_2$ together provide complete information about outcomes $S = 0$ and $S = 2$, which is not true for either alone. Thus, in this case interaction information is zero despite the presence of both redundant and synergistic interactions, because redundancy and synergy are balanced.

The situation is worse for four-variable interaction information, which is known to violate the interpretation that positive values indicate (pure) synergy and negative values indicate (pure) redundancy [5,21]. To demonstrate, consider the case of 3-parity, which is the higher-order form of the exclusive-OR, or 2-parity, function mentioned earlier. In this case, we have a system of four binary random variables, $S$ and $\mathbf{R} = \{R_1, R_2, R_3\}$, where the eight outcomes for $\mathbf{R}$ are equiprobable and $S = R_1 \oplus R_2 \oplus R_3$. Intuitively, this corresponds to a case of pure synergy, since the value of $S$ can be determined only when all of the $R_i$ are known. Indeed, using Equation (4.18) we find that $I(S; R_1; R_2; R_3)$ is equal to $+1$ bit, as expected from the interpretation that positive values indicate synergy. However, now consider a second system of binary variables, this time where the two outcomes of $S$ are equiprobable and $R_1$, $R_2$, and $R_3$ are all copies of $S$. Clearly this corresponds to pure redundancy, since the value of $S$ can be determined uniquely from any $R_i$, but $I(S; R_1; R_2; R_3)$ for this system is again equal to $+1$ bit, same as the case of pure synergy. Thus, a completely redundant system is assigned a positive value for the interaction information, clearly violating the idea that redundancy is indicated by negative values. Worse still, the 4-variable interaction information fails to distinguish between the polar opposites of purely synergistic and purely redundant information.

The PI-decomposition for 4-variable interaction information (Figure 4.9) clarifies why this is the case. From Equations (4.10) and (4.18), it follows that $I(S;R_1;R_2;R_3)$ is equal to

$$\begin{aligned}
&\Pi(S;\{123\}) + \Pi(S;\{1\}\{2\}\{3\}) \\
&-\Pi(S;\{1\}\{23\}) - \Pi(S;\{2\}\{13\}) - \Pi(S;\{3\}\{12\}) \\
&-\Pi(S;\{12\}\{13\}) - \Pi(S;\{12\}\{23\}) - \Pi(S;\{13\}\{23\}) \\
&-2 \times \Pi(S;\{12\}\{13\}\{23\}). \qquad (4.20)
\end{aligned}$$

Thus, $I(S;R_1;R_2;R_3)$ equals the sum of third-order synergy ({123}) and third-order redundancy ({1}{2}{3}), minus the information provided redundantly by any first- and second-order synergy ({1}{23}, {2}{13}, and {3}{12}), minus the information provided redundantly by any two second-order synergies ({12}{13}, {12}{23}, and {13}{23}), and minus twice the information provided redundantly by all three second-order synergies ({12}{13}{23}). As a result, systems with pure synergy and pure redundancy have the same value for $I(S;R_1;R_2;R_3)$ because 4-variable interaction information adds in the highest-order synergy and redundancy terms. More generally, the PI-decomposition shows why $I(S;R_1;R_2;R_3)$ is difficult to interpret as a meaningful quantity, and as one might expect the story only becomes more complicated in higher dimensions. Thus, although one can readily decompose interaction information into a collection of partial information contributions, and understand the conditions under which it will be positive or negative depending on the relative magnitudes of these contributions, the utility of interaction information for higher-dimensional systems is unclear.

## 4.7 Discussion

In summary, the main objective of this chapter has been to quantify multivariate information in a way that illuminates the structure of variable interactions. This was accomplished by first defining an axiomatic measure of redundancy, $I_\cap$, that captures the intuition of redundancy as overlapping information. Next, it was shown that $I_\cap$ induces a lattice over the set of possible information sources, referred to as the redundancy lattice, which characterizes the distinct ways that information can be distributed amongst a set of sources. From this lattice, a measure of partial information was derived that captures the unique information contributed by each possible combination of sources. It was then shown that mutual information decomposes into a sum of these partial information terms, so that the total information provided by a source is broken down into a collection of partial information contributions. Moreover, it was demonstrated that each of these terms supports a clear interpretation as a particular combination of redundant and synergistic interactions between specific subsets of variables. Finally, we discussed the relationship between PI-decomposition and interaction information, the current de facto measure of multivariate information, and used partial information to clarify the confusing property that interaction information is sometimes negative.

One obvious challenge with applying these ideas is that the number of PI-terms grows rapidly for larger systems. For instance, with 9 variables there are more than $5 \times 10^{22}$ possibilities [248], and beyond that the Dedekind numbers are not even currently known. Thus, clearly an important direction for future work is to determine efficient ways of calculating partial information for larger systems. To this end, the lattice structure of the terms is likely to play an essential role. As with any ordered data structure, the fact that the space of possibilities is highly organized can be readily exploited for efficient use. For instance, as a simple example, if $I_{\min}$ is calculated in a descending fashion over the nodes of the redundancy lattice and at a certain juncture has a value of zero, all of the terms below that node can immediately be eliminated simply from the monotonicity of $I_\cap$. Moreover, if the Markov property or any other constraints hold between the variables, many of the possible PI-terms can also be excluded. Finally, these considerations notwithstanding, it should also be emphasized that 3-variable interaction is the current state of the art, and thus even the simplest form of PI-decomposition can be used to address a number of outstanding questions.

In physics, for example, 3-variable interactions have been explored in relation to the non-separability of quantum systems [29] and in the study of many-body correlation effects [151]. In neuroscience, the concepts of synergy and redundancy for three variables have been examined in the context of neural coding in a number of theoretical and empirical investigations [25, 68, 131, 158, 169, 196]. In genetics, multivariate dependencies arise in the analysis of gene-gene and gene-environment interactions in studies of human disease susceptibility [5, 30, 156]. Similar issues have also been explored in machine learning [108, 146, 172], ecology [164], quantum information theory [240], information geometry [2], rough set analysis [70], and cooperative game theory [82]. In all of these cases, the 3-variable form of PI-decomposition can be applied immediately to illuminate the structure of multivariate dependencies, while the general form provides a clear way forward in the study of more complex systems of interactions.

# 5

# A Toolkit for Information Dynamics

In this chapter, we develop techniques for quantifying information dynamics using the method of PI-decomposition described in the previous chapter. First, in Section 5.1, we develop techniques for quantifying the flow of *intrinsic* information, or the flow of information within a stochastic process $X$ where the informational quantity of interest is a property of $X$ itself. The canonical measure of this kind is the transfer entropy (Section 2.7), which quantifies the transfer of information into a process $X$ that is about the future state of $X$. Indeed, the techniques developed in Section 5.1 are best thought of as extensions of transfer entropy, allowing one to tease apart intrinsic information flow in ways that are not possible using transfer entropy alone. Second, in Section 5.2, we develop techniques for quantifying the flow of *extrinsic* information, or the flow of information within and between stochastic processes that is about something external to those processes. These techniques will be put into action in Chapter 7, where we will use them to explore how information about external stimulus features flows through the components of brain-body-environment systems. Finally, in Section 5.3, we conclude by discussing the relationship between measures of intrinsic and extrinsic information flow and identifying some promising directions for future work.

## 5.1 Quantifying the Flow of Intrinsic Information

As discussed in Section 2.7, transfer entropy [113,198] provides a directional measure of the influence that one random process, the *source*, has on another, the *target*. This influence is measured by the information that the source provides about the next state of the target when conditioned on the target's history. Transfer entropy has become widely adopted as a standard measure of information transfer, with applications in neuroscience [80,98,218], cellular biology [166], chaotic synchronization [105,165,167], and econophysics [130,149] to name just a few. The idea behind transfer entropy is that conditioning on the target's history removes the information shared by the source and target due to common histories or inputs, thereby isolating the information that is actually transferred. However, conditioning does not simply remove shared information; it also adds in higher-order synergistic information, an idea that can be formalized using PI-decomposition.

In this section, we apply this basic property of conditional information to generalize transfer entropy in two complementary ways. First, we decompose transfer entropy into two kinds of information transfer that differ regarding the influence of the target's state. We show that the resulting measures are formally related to the control-theoretic concepts of open-loop and closed-loop control, and quantify separately the state-independent and state-dependent influences of the source onto the target. Second, we apply a similar decomposition to the case of multiple sources and derive a novel multivariate generalization of transfer entropy. The resulting measures quantify separately the unique, redundant, and synergistic influences of multiple sources onto a target. Together these results provide a general framework for characterizing not only the magnitudes and directions—but also the *kinds*—of information exchange that occur between random processes.

We begin by considering the PI-decomposition for the conditional mutual information

$I(X;Y|Z)$ between three random variables $X$, $Y$, and $Z$. Using the chain rule for information (Equation (2.19)), $I(X;Y|Z)$ can be written as

$$I(X;Y|Z) = I(X;Y,Z) - I(X;Z),$$

and expanding into PI-terms (Equations (4.12) and (4.13)),

$$I(X;Y|Z) = \Pi(X;\{Y\}) + \Pi(X;\{Y,Z\}). \tag{5.1}$$

In other words, $I(X;Y|Z)$ decomposes into the unique information from $Y$ *plus* the synergy from $Y$ and $Z$. Thus, conditioning $I(X;Y)$ on $Z$ not only removes the redundancy from $Y$ and $Z$, but also adds in their synergy. This observation makes intuitive sense if we think of $I(X;Y|Z)$ as answering the question: How much information do we gain from learning $Y$ when we already know $Z$? Clearly, this will include both the information that comes uniquely from $Y$ plus the synergistic information that comes from $Y$ and $Z$ together.

Transfer entropy can be analyzed in a similar way, since it is simply an application of conditional mutual information to stochastic processes. As discussed in Section 2.7, for stochastic processes $X$ and $Y$ the transfer entropy from $Y$ to $X$ is defined as

$$T_{Y\to X} = I(X_{t+1}; Y_t^{(l)}|X_t^{(k)}), \tag{5.2}$$

where $X_t^{(k)}$ is the $k$-dimensional delay vector for $X$, and likewise for $Y_t^{(l)}$ and $Y$ (henceforth the superscripts are omitted for clarity). In other words, $T_{Y\to X}$ quantifies the information that previous values of $Y$ provide about the next state of $X$ when conditioned on $X$'s own history. In terms of transition probabilities, $T_{Y\to X}$ can also be thought of as quantifying deviation from the generalized Markov property

$$p(x_{t+1}|x_t, y_t) = p(x_{t+1}|x_t),$$

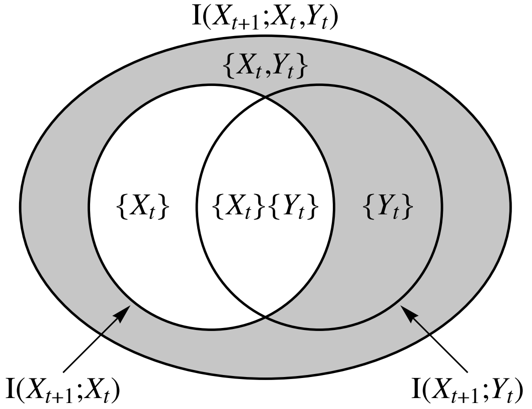


Figure 5.1: PI-decomposition of transfer entropy. $I(X_{t+1}; X_t, Y_t)$ decomposes into unique information from $X_t$ ($\{X_t\}$) and $Y_t$ ($\{Y_t\}$), redundancy ($\{X_t\}\{Y_t\}$), and synergy ($\{X_t, Y_t\}$). The transfer entropy $T_{Y\to X}$ corresponds to the gray region, which decomposes into SITE ($\{Y_t\}$) and SDTE ($\{X_t, Y_t\}$).

with $T_{Y\to X} = 0$ if, and only if, $Y$ has no influence on the state transitions of $X$.

Our first main result of this section is that, by decomposing $T_{Y\to X}$, we can distinguish two kinds of information transfer, or two distinct ways that $Y$ can influence the transitions of $X$ (Figure 5.1). Letting $\mathbf{R} = \{X_t, Y_t\}$ and combining Equations (5.1) and (5.2), we have that

$$T_{Y\to X} = \Pi(X_{t+1}; \{Y_t\}) + \Pi(X_{i+1}; \{X_t, Y_t\}) \tag{5.3}$$

where the first PI-term is the unique information that $Y_t$ provides about $X_{t+1}$ and the second PI-term is the synergistic information from $X_t$ and $Y_t$. As we will show, $\Pi(X_{t+1}; \{Y_t\})$ corresponds to *state-independent transfer entropy* (SITE): it measures the portion of $Y_t$'s influence on $X_{t+1}$ that does not depend on $X_t$. The complementary term $\Pi(X_{t+1}; \{X_t, Y_t\})$ is the *state-dependent transfer entropy* (SDTE): it measures the influence that $Y_t$ has on $X_{t+1}$ only when combined with an appropriate state of $X_t$. To ground this interpretation, we next establish a formal connection between SITE and SDTE and the control-theoretic notions of open-loop and closed-loop control.

In control theory, one considers a process $X_t$—characterized by its initial state $X$ and final state $X'$—and a controller $C$, with the two related by a distribution $p(x'|x,c)$ that specifies the probability of transitioning from an initial state $x$ to a final state $x'$ under control action $c$ [234, 235]. When designing a controller, the aim is to specify a control policy, given by the distribution $p(c|x)$, that moves the system to certain desired final states. In open-loop control, the controller $C$ is forced to act independently of the initial state $X$ (i.e., $I(X;C) = 0$), while closed-loop control is characterized by state-dependent actuation.

A fundamental property of a control system is its *controllability*, or the extent to which the controlled process can be moved through its entire state space. In particular, a system is said to have perfect controllability if, and only if, there exists a control policy that can move the system deterministically from any initial state $x \in X$ to any final state $x' \in X'$. In [235], it is shown that a natural information-theoretic measure of controllability is $I(X';C|X)$—the *information transfer* from the controller to the controlled process—which takes on its maximum value exactly in the case of perfect controllability. Thus, there is a close parallel between information transfer and controllability, where essentially the only difference is semantic: information transfer applies to arbitrary interactions between processes, while controllability is concerned specifically with using one process to control another.

With this in mind, the following result connects SITE and SDTE with open-loop and closed-loop control[1].

**Theorem 5.1.** *A system is perfectly controllable with open-loop control iff it is perfectly controllable with only state-independent transfer from $C$ to $X'$.*

Thus, decomposing $I(X';C|X)$ as in Equation (5.3), SITE from $C$ to $X'$ measures a system's open-loop controllability (maximal for perfect open-loop control), while SDTE measures the additional contribution from closed-loop control. More generally, this connection grounds the interpretation of SITE as the state-independent (open-loop) influence of one process on another, and likewise for SDTE and state-dependent (closed-loop) influence.

[1] The proof of this result is given in Appendix B.

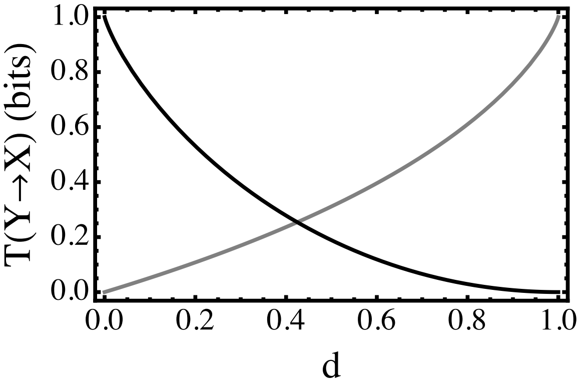


Figure 5.2: Example of SITE and SDTE. The SITE (black) and SDTE (gray) for binary Markov processes $X$ and $Y$ are plotted as a function of the coupling parameter $d$.

As a simple example to illustrate the two kinds of transfer, consider two binary state Markov processes $X$ and $Y$, where $Y$ is purely random and $X$ is stochastically coupled to $Y$. Specifically, if $x_t = 0$, then $x_{t+1} = y_t$, while if $x_t = 1$, the probability that $x_{t+1} = y_t$ is $1 - d$ and that $x_{t+1} = 1 - y_t$ is $d$. A simple eigenvector calculation yields the stationary distribution $p(x, y) = \frac{1}{4}$ for all $x$ and $y$, and from this all informational quantities can be computed. When $d = 0$, $X_{t+1}$ is simply set to $Y_t$ regardless of its own previous state[2], thus corresponding to pure SITE (Figure 5.2). In contrast, when $d = 1$, $y_t = 0$ causes $X$ to remain in the same state and $y_t = 1$ causes $X$ to switch states. Consequently, $Y$'s influence on $X_{t+1}$ depends entirely on $X_t$, corresponding to pure SDTE. In fact, if one imagines using $Y$ to control $X$, then $d = 1$ corresponds to a 'controlled-NOT' gate, a classic example of a system that requires closed-loop control [234, 235]. Figure 5.2 shows how varying $d$ produces a smooth transition between these two extremes.

The distinction between SITE and SDTE also clarifies the relationship between transfer entropy and the time-delayed mutual information (TDMI) $I(X_{t+1}; Y_t)$, which was the standard measure of information transfer prior to transfer entropy [198]. Transfer entropy was initially proposed as an alternative to TDMI because the latter fails to remove shared

[2] With this parameter setting, the system is equivalent to the discrete example considered in [113].

information due to common histories or inputs. From Figure 5.1, it is clear that this shared information corresponds to $\Pi(X_{t+1};\{X_t\}\{Y_t\})$, the redundancy between $X_t$ and $Y_t$. However, Figure 5.1 also reveals a second crucial difference between TDMI and transfer entropy, which is that TDMI fails to include SDTE. Thus, not only does TDMI incorrectly add in shared information, but it also leaves out a significant component of information transfer.

Our second main result of this section is a novel multivariate generalization of transfer entropy, based on applying PI-decomposition to the information from multiple sources. Schreiber [198] originally proposed a generalization of transfer entropy based on 'conditioning out' other sources, an idea that has since been adopted and extended by others [63, 139]. However, it should be clear from the preceding discussion that such a generalization is problematic, since conditioning on other sources does not simply remove their shared information. Our generalization addresses this deficiency by quantifying separately the unique, redundant, and synergistic transfer from multiple sources.

For simplicity, we consider only two sources $Y$ and $Z$ acting on a target $X$ (the general case is discussed momentarily), in which case the total transfer entropy is given by $T_{Y,Z\to X} = I(X_{t+1};Y_t,Z_t|X_t)$. Applying PI-decomposition as before, we arrive at measures for the *redundant transfer* from $Y$ and $Z$: $T_{\{Y\}\{Z\}\to X} = I_{\min}(X_{t+1};Y_t,Z_t|X_t)$; the *unique transfer* from Y (resp. Z): $T_{Y\to X\setminus Z} = T_{Y\to X} - T_{\{Y\}\{Z\}\to X}$; and the *synergistic transfer* from $Y$ and $Z$: $T_{\{Y,Z\}\to X} = T_{Y,Z\to X} - I_{\max}(X_{t+1};Y_t,Z_t|X_t)$. Generally speaking, redundant transfer corresponds to situations where the apparent influence from multiple sources may in fact be due to any one (or several) of them, indicating that interventional methods are required to determine the true causal structure [10]. In contrast, unique transfer represents the portion of a source's influence that can only come from that source and not the others, or, if all possible sources are considered, the portion that *must* come from that source.

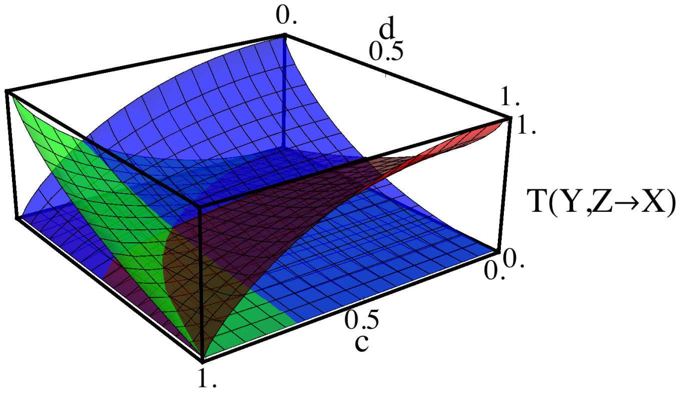


Figure 5.3: Example of multivariate transfer entropy. $T_{Y\to X\setminus Z}$ (blue), $T_{\{Y\}\{Z\}\to X}$ (green), and $T_{\{Y,Z\}\to X}$ (red) for binary Markov processes $X$, $Y$, and $Z$ are plotted as a function of the coupling parameters $c$ and $d$.

Finally, synergistic transfer indicates that several sources act together cooperatively to influence the target.

To illustrate, consider three binary state Markov processes $X$, $Y$, and $Z$. $Y$ is purely random, and $Z$ is stochastically coupled to $Y$ such that $z_t = y_t$ with probability $(1+c)/2$ and $z_t = 1 - y_t$ with probability $(1-c)/2$. This coupling can be thought of as an external signal driving $Y$ and $Z$ to synchronize: as $c$ goes from $0$ to $1$, $Y$ and $Z$ transition from independence to complete synchronization. $X$ in turn is coupled to both $Y$ and $Z$ such that, if $z_t = 0$, $x_{t+1} = y_t$, while if $z_t = 1$, $x_{t+1} = y_t$ with probability $(1-d)$ and $x_{t+1} = (1-y_t)$ with probability $d$. Thus, $x_{t+1} = y_t$ when $d = 0$, and $x_{t+1} = y_t \oplus z_t$ when $d = 1$. At the extreme parameter settings, this system exhibits three different behaviors (Figure 5.3). With $(c = 0, d = 0)$, $Y$ and $Z$ are independent and $X$ depends only on $Y$, so the only influence is unique transfer from $Y$ to $X$. In contrast, with $(c = 1, d = 0)$, $X$ again depends only on $Y$ but $Y$ and $Z$ are now synchronized, so there is only redundant transfer from $Y$ and $Z$. Indeed, in this case it is impossible to determine from observations alone whether $Y$ or $Z$ (or both) is driving $X$. Finally, with $(c = 0, d = 1)$, $Y$ and $Z$ are independent and $X_{t+1} = Y_t \oplus Z_t$, corresponding to pure synergistic transfer.

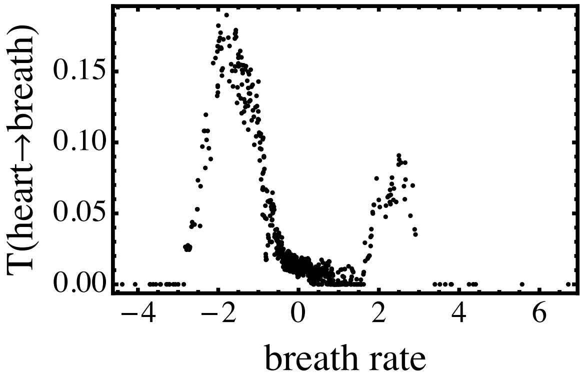


Figure 5.4: SDTE from heart rate to breath rate. The information transfer $I(B_{t+1}; H_t | B_t = b_t)$ from heart rate ($H$) to breath rate ($B$) is plotted as a function of $b_t$ for bandwidth $r = 0.5$. Qualitatively similar results were found for $r \in [0.2, 1.0]$.

As a final example, we extend the analysis of a multivariate physiological time series presented in [113, 198]. The data consists of simultaneous recordings of the breath rate (chest volume), heart rate, and blood oxygen concentration for a patient suffering from sleep apnea. Previous analysis compared transfer entropy and TDMI for both directions between the breath and heart signals. However, directly comparing transfer entropy and TDMI is problematic and potentially misleading, since both measures detect SITE but differ regarding SDTE and shared information (Figure 5.1). Indeed, with no additional information, it is impossible to determine even whether transfer entropy and TDMI are detecting the same or different aspects of an interaction.

To address this issue, we calculated SITE and SDTE between the breath and heart signals. Joint probability estimates were obtained by kernel estimation using a rectangular kernel with bandwidth $r$. Neighboring points closer than 20 time steps were excluded and points with fewer than 5 neighbors were ignored, following the suggestions of Schreiber and others [113,166,198]. Our main finding is that SITE is consistent with zero in both directions between the breath and heart signals. Thus, for these signals, transfer entropy and TDMI in fact quantify entirely separate things: TDMI is due only to shared information

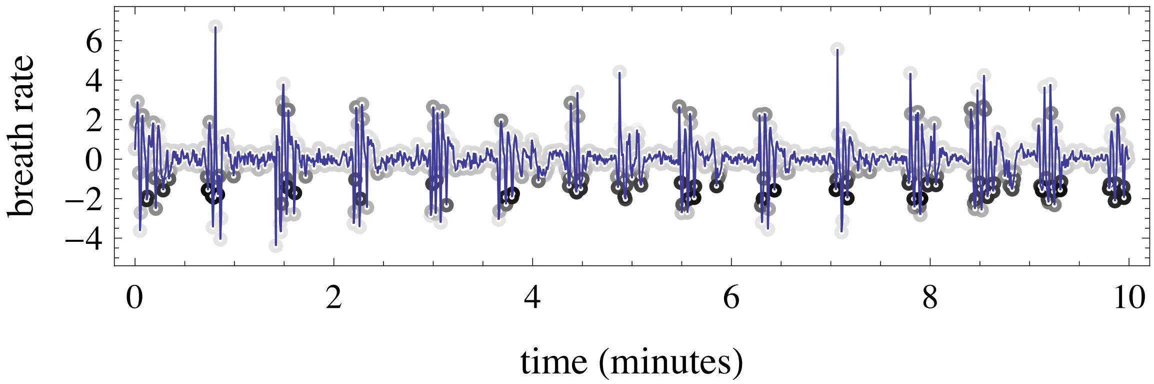


Figure 5.5: SDTE superimposed on the breath rate signal. Darker colors correspond to higher values of $I(B_{t+1}; H_t | B_t = b_t)$. Heart rate has the largest influence on breath rate when chest volume is low and, to a lesser extent, when it is high, but minimal influence when chest volume is near its mean.

from common histories or inputs, while transfer entropy detects state-dependent information exchange. This state dependence can be seen by plotting information transfer as a function of the target state, shown in Figure 5.4 for $T(\text{heart} \to \text{breath})$. For pure SITE, this plot would be uniform across target states, while Figure 5.4 shows a clear bimodal distribution. These two modes correspond to downswings and upswings in chest volume, suggesting that heart rate has the largest influence on respiration when chest volume is low and, to a lesser extent, when it is high, but minimal influence when chest volume is near its mean. This result is illustrated most clearly in Figure 5.5, where the information transfer values are superimposed on the breath rate signal.

Finally, we also analyzed the combined influence of heart rate and blood oxygen level on breath rate (Figure 5.6). We found that the most significant component is consistently the unique information transfer from the heart rate, indicating that most of the transfer entropy discussed above is uniquely attributable to the heart signal. However, there is also considerable redundant and synergistic transfer, of roughly comparable magnitude, from heart rate and blood oxygen concentration. In contrast, there is essentially no unique information transfer from blood oxygen concentration, indicating that all of its apparent

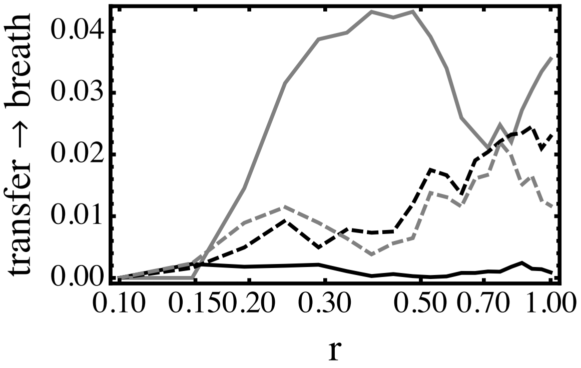


Figure 5.6: Multivariate transfer for the physiological time series. $T_{O \to B \setminus H}$ (black), $T_{H \to B \setminus O}$ (gray), $T_{\{H\}\{O\} \to B}$ (dashed black), and $T_{\{H,O\} \to B}$ (dashed gray) from heart rate (H) and blood oxygen (O) to breath rate (B) are plotted as a function of the bandwidth $r$.

influence could also be due to heart rate.

## 5.2 Quantifying the Flow of Extrinsic Information

In this section, we develop a complementary set of measures for quantifying the flow of extrinsic information, or information about some variable that is external to the source and target processes. For example, in a neuroscience context, the external variable could represent a stimulus ensemble while the source and target processes represent the responses of different neurons. In this case, our measures would quantify the information about the stimulus that is carried by and exchanged between the neurons. The ideas behind these extrinsic measures follow a parallel line of reasoning to those underlying transfer entropy, the canonical measure of intrinsic information transfer. Thus, for expository purposes, we will develop our measures of extrinsic information flow by tracing the parallels with transfer entropy. In particular, we will show that both transfer entropy and our extrinsic measures follow from identifying: (1) an informational quantity of interest associated with the target process; (2) how that quantity changes from one moment to the next; and (3) what portion of that change can be accounted for by the source process.

For transfer entropy, the informational quantity of interest is the entropy $H(X_{t+1})$. That is, for transfer entropy, the central aim is to account for changes in $H(X_{t+1})$; specifically, the source $Y$ is said to transfer information to $X$ if it reduces $H(X_{t+1})$, i.e., if it reduces uncertainty regarding $X$'s next state. Note that this is the sense in which transfer entropy is an intrinsic measure, as the informational quantity of interest is a property of the target process itself. In contrast, for our measures of extrinsic information flow, the informational quantity of interest is $I(F; X_{t+1})$, where $F$ is an arbitrary random variable assumed to reflect some property of the external environment. In other words, we want our measures to account for changes in the information that the process $X$ provides *about* $F$. For both transfer entropy and our extrinsic measures, the quantities of interest can be depicted diagrammatically as in panels A1 and B1 of Figure 5.7. Note that these diagrams are the simplest possible forms for I-diagrams (Figure 2.1) and PI-diagrams (Figure 4.5), respectively. Indeed, throughout this section the parallels between transfer entropy and our extrinsic measures will be mirrored by analogous parallels between I-diagrams and PI-diagrams.

The next step is to identify how the quantities of interest change over time. In particular, we want to isolate the portion of each informational quantity that is contained by $X$ at time $t + 1$ but was not contained by $X$ at time $t$. For transfer entropy, this change in the quantity of interest is given by the conditional entropy $H(X_{t+1}|X_t)$, or the uncertainty regarding $X_{t+1}$ that is not accounted for by $X_t$. In dynamic terms, this quantity corresponds to the *entropy gain* for the process $X$. Note that the complementary portion of $H(X_{t+1})$, the portion that is contained by $X$ at time $t$, is given by the mutual information $I(X_{t+1}; X_t)$. Again, the relationship between these quantities can be represented with an I-diagram (panel A2 of Figure 5.7), where $H(X_{t+1}|X_t)$ corresponds to the set difference $H(X_{t+1}) \setminus H(X_t)$ and $I(X_{t+1}; X_t)$ represents the intersection of $H(X_t)$ and $H(X_{t+1})$.

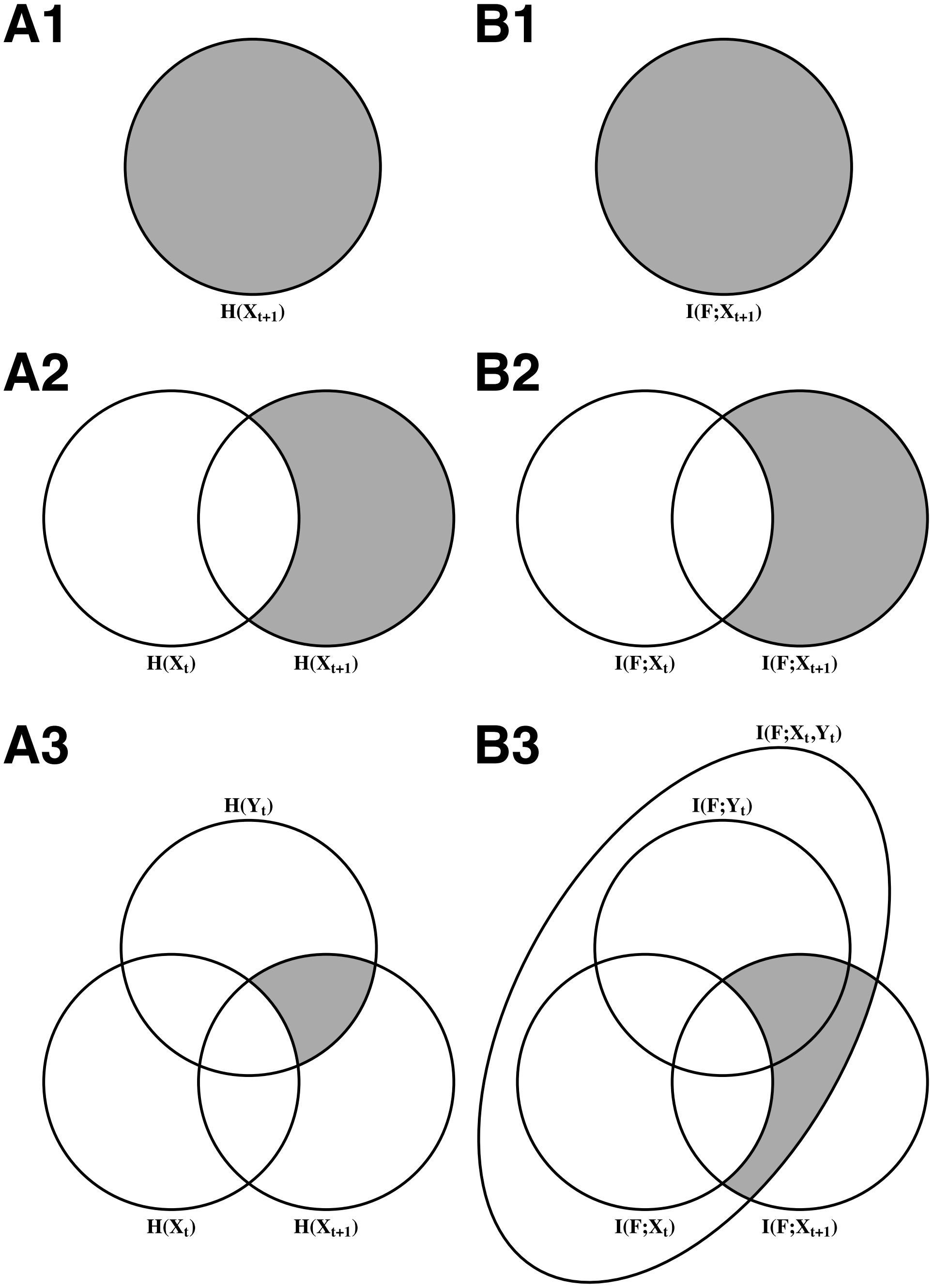


Figure 5.7: Parallel development of intrinsic and extrinsic flow measures. For intrinsic flow, I-diagrams depict the informational quantity of interest (A1), the change in that quantity (A2), and the transfer of that quantity (A3). PI-diagrams representing the analogous quantities for extrinsic flow are shown in B1-B3.

In contrast, for extrinsic information flow, the intersection of $I(F;X_t)$ and $I(F;X_{t+1})$—our quantity of interest at the current and next time steps—is given by the redundancy $I_{\min}(F;X_{t+1},X_t)$. Consequently, the *information gain* at time $t+1$ is given by

$$I_G(F;X_{t+1}) = I(F;X_{t+1}) - I_{\min}(F;X_{t+1},X_t), \tag{5.4}$$

or the unique information that $X$ provides about $F$ at time $t+1$[3]. Diagrammatically, these terms have exactly the same structure as those corresponding to measures of intrinsic information (compare panels A2 and B2 of Figure 5.7), though with a PI-diagram in place of an I-diagram.

The final step is to identify the portion of the new entropy/information associated with $X_{t+1}$ that can be accounted for by the source process $Y$, indicating potential information transfer from $Y$ to $X$. In other words, we need to answer the question: Of the information gained by $X$ at time $t+1$, how much was shared by, and thus was potentially transferred from, $Y$ at time $t$? Of course, as the discussion in the preceding section hopefully made clear, we need to consider not only the information shared by $Y_t$, but also the information shared by the synergy $\{X_t,Y_t\}$, thus accounting for the possibility of synergistic (state-dependent) transfer. For transfer entropy, the portion of $H(X_{t+1})$ that is shared by $X_t$, $Y_t$, or the synergy $\{X_t,Y_t\}$, is given by the mutual information $I(X_{t+1};X_t,Y_t)$. Thus, subtracting from this quantity the mutual information $I(X_{t+1};X_t)$—the portion of $H(X_{t+1})$ that is shared by $X_t$—yields the portion of $H(X_{t+1})$that is shared by $Y_t$ and $\{X_t,Y_t\}$, the transfer

[3] It is worth noting that, complementary to this measure of information gain, one can also define a measure of information loss $I_L(F;X_{t+1})$ by simply reversing $X_{t+1}$ and $X_t$ in Equation (5.4). This measure quantifies the information that $X$ contained about $F$ at time $t$ that it lacks at time $t+1$. Similarly, one can define entropy loss as $H(X_t|X_{t+1})$. Taken together, information gain and loss specify how the information contained by a component changes from one moment to the next:

$$I(F;X_{t+1}) = I(F;X_t) - I_L(F;X_{t+1}) + I_G(F;X_{t+1}).$$

However, while information loss happens generically for the components of a dissipative dynamical system, information gain signals the influence of another component in transferring information, which is the reason why we focus only on information gain here.

entropy $T_{Y\to X}$ (panel A3 of Figure 5.7):

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

Analogously, for extrinsic information flow, the portion of $I(F;X_{t+1})$ that is shared by $X_t$, $Y_t$, or $\{X_t,Y_t\}$, is given by the redundancy $I_{\min}(F;X_{t+1},\{X_t,Y_t\})$. Thus, subtracting from this quantity the redundancy $I_{\min}(F;X_{t+1},X_t)$—the portion of $I(F;X_{t+1})$ that is shared by $X_t$—yields the portion of $I(F;X_{t+1})$ that is shared by $Y_t$ and $\{X_t,Y_t\}$ (panel B3 of Figure 5.7). We will call this measure the information transfer[4] from $Y$ to $X$ *about* $F$:

$$T(F;Y\to X) = I_{\min}(F;X_{t+1},\{X_t,Y_t\}) - I_{\min}(F;X_{t+1},X_t). \tag{5.5}$$

Lastly, as a sanity check, we can verify that our extrinsic and intrinsic measures are equivalent in the case that $F = X_{t+1}$, i.e., when the "extrinsic" variable $F$ is equivalent to the intrinsic variable representing the next state of $X$. First, in this case the quantity of interest $I(F;X_{t+1})$ simplifies to $I(X_{t+1};X_{t+1}) = H(X_{t+1})$, or the self-information for $X_{t+1}$. Second, the information gain $I_G(F;X_{t+1})$ can be rewritten as

$$\begin{aligned}I_G(X_{t+1};X_{t+1}) &= I(X_{t+1};X_{t+1}) - I_{\min}(X_{t+1};X_{t+1},X_t)\\ &= H(X_{t+1}) - I(X_{t+1};X_t)\\ &= H(X_{t+1}|X_t),\end{aligned}$$

where the second step follows from Theorem 4.3, since $X_{t+1}\to X_{t+1}\to X_t$. Third and

[4] Note that, as with the decomposition of $T_{Y\to X}$ into SDTE and SITE, one could also decompose $T(F;Y\to X)$ into separate contributions from $Y_t$ and $\{X_t,Y_t\}$ (the two PI-terms shaded in panel B3 of Figure 5.7), corresponding to information transferred uniquely from $Y$ versus information transferred synergistically by $X$ and $Y$. However, we will ignore this distinction here and instead treat $T(F;Y\to X)$ as a unitary measure.

finally, the information transfer $T(F; Y \to X)$ simplifies as follows:

$$\begin{aligned} T(X_{t+1}; Y \to X) &= I_{\min}(X_{t+1}; X_{t+1}, \{X_t, Y_t\}) - I_{\min}(X_{t+1}; X_{t+1}, X_t) \\ &= I(X_{t+1}; X_t, Y_t) - I(X_{t+1}; X_t) \\ &= I(X_{t+1}; Y_t | X_t) = T_{Y \to X}. \end{aligned}$$

Thus, our “extrinsic” measures of information flow may be more aptly named *generalized* measures of information flow, as they subsume the intrinsic measures in the special case that $F$, an arbitrary variable, corresponds to the next state of the target process itself. However, we retain the label ”extrinsic” here in light of our later applications to brain-body-environment systems, where these measures will be used to quantify the flow of information, through neural and bodily variables, about features of the external world.

## 5.3 Discussion

The fact that transfer entropy, or intrinsic information flow more generally, can be derived as a special case of our ”extrinsic” (or generalized) measures suggests a number of intriguing questions for future work. For instance, is there a useful physical and/or computational distinction between one system transferring entropy (increasing predicability) in another system versus one system transferring (potentially useful) information to another? If so, in what sense is the former a special case of the latter (since intrinsic measures correspond to a single parameter setting—$F = X_{t+1}$—of extrinsic ones)? Or, more generally, what is the relationship between these two classes of measures? One interesting possibility is that intrinsic information flow may correspond to the upper bound on extrinsic information flow. This idea seems mathematically plausible considering that mutual information is maximally equal to the entropy ($I(F; X_{t+1}) \leq H(X_{t+t})$), with the latter commonly interpreted as a variable’s “maximum information-carrying capacity”.

The idea also seems intuitively reasonable, as the total influence that $Y$ has on $X$ is surely an upper bound on the information that $Y$ can transfer to $X$ about some variable $F$. If this relationship proves to be true it could, for example, have useful applications in fields like neuroscience, where measures of intrinsic flow could first be used to identify neurons that are significant transmitters and/or receivers of information, and subsequently measures of extrinsic flow could be used to tease apart the specific messages carried by their incoming and outgoing transmissions.

On a related note, another interesting possibility for future work involves the measure of information loss, complementary to that of information gain, mentioned in the previous section. In particular, by following the same steps that led from information gain to information transfer, one could similarly derive a measure of the transfer of information loss. In other words, while our measure of information transfer quantifies the extent to which one process anticipates (or drives/causes) information gain in another process, one could similarly quantify the extent to which one process leads to information loss in another. What might be the computational, or cognitive, significance of the transfer of information loss? The precise answer is unclear, but one could envision such transfer playing a crucial role in, e.g., attentional processes, where it is essential to ignore, or to systematically lose, information about extraneous features.

Finally, we conclude by mentioning how the measures presented in this chapter generalize to larger systems of interacting processes. First, note that the multivariate generalization of transfer entropy described in Section 5.1 extends naturally to any number of information sources simply by applying the general form of PI-decomposition. The two extensions of transfer entropy described in Section 5.1 can also be applied in conjunction, allowing one to quantify, e.g., state-dependent synergistic transfer. Finally, the extrinsic measures introduced in Section 5.2 are also easily extended using higher-order PI-decompositions, so that one can quantify how multiple processes interact to carry and transfer information about external features. Thus, together these techniques provide a completely general framework for characterizing the dynamics of information exchange in complex systems.

# 6

# Embodied Models of Relational Categorization

This chapter introduces the model agents that will be used in Chapter 7 to illustrate the application of information dynamics to embodied cognitive systems. These agents perform a simple kind of relational categorization, which is a phenomenon that has garnered considerable interest in cognitive science and poses a number of challenging problems for analysis. Two key properties of these model agents distinguish them from other models of relational categorization and make them ideally suited for our purposes here. First, with these models, relational categorization is defined as a particular kind of behavior exhibited by embodied agents, rather than as an abstract computational task as it has traditionally been studied. Second, the models were developed using an evolutionary paradigm that imposes minimal constraints on how the relational behavior is implemented. In particular, the evolved solutions are free to span the brain-body and body-environment boundaries, making full use of the resources available to an embodied agent. Thus, from an analytical perspective, these models present us with many of the same challenges as organisms in the real world, for which cognition is typically identified as a set of adaptive behaviors and it is the task of analysis to reveal the mechanisms underlying these behaviors.

The chapter begins with a brief overview of relational categorization and its importance for cognitive science. Then we describe previous efforts to model relational categorization and elaborate on the important features that distinguish our approach. Afterwards, we detail the methods used to evolve our model agents and present the results from a set of evolutionary studies. This is followed by an analysis of the best evolved agent using the mathematical tools of dynamical systems theory, with this analysis providing some preliminary insights into the mechanisms underlying relational categorization in our model agents. Finally, we conclude the chapter by summarizing the main findings from our dynamical analysis, thereby setting the stage for further exploration using the techniques of information dynamics in Chapter 7.

## 6.1 Relational Categorization

Relations and relational categories have attracted a considerable amount of attention in cognitive science and related fields [72, 78, 90, 116, 129, 148]. Relational categories are categories that are determined by common relational structure among category members, in contrast with object categories, which are determined by intrinsic similarities between members. Thus, relational categories, such as *same* or *smaller*, can include instances that have few or no intrinsic similarities (e.g., Texas is *smaller* than Alaska; Muggsy Bogues is *smaller* than Shaquille O'Neal) so long as they share the core relationship, whereas object categories, such as *round* or *camel*, are confined to instances that all share intrinsic properties (e.g., the properties of *roundness* or *camelness*). Because of this, relational categorization is often considered a hallmark of "higher-level" cognition, since it requires the ability to abstract properties of individual entities in order to identify their higher-order similarities. Indeed, relational categories are fundamental to many topics in higher cognition, such as analogy and metaphor, language, and mathematics [71, 90]. However, relational categorization is not an ability that is restricted to humans and other higher organisms. Rather, a sensitivity to relational categories has been demonstrated in a wide range of species, including pigeons [35, 254, 260], baboons [58, 246], rats [163, 194], and insects [75, 154].

Relational categorization tasks can be described in terms of two sets of features of the related objects, commonly referred to as *roles* and *fillers*. To illustrate this idea, consider a situation in which a book is above a table. Each object has both an object category (*book*, *table*) as well as a spatial relational feature (*above*, *below*). A crucial requirement for relational categorization is that one is able to distinguish this situation from the opposite situation in which the table is above the book. To do so, one must associate or bind each object category with its appropriate relational feature, where the latter can be thought of as roles and the former as fillers for those roles. Thus, *book* must be bound to the *above* role and *table* to the *below* role. In the relational task explored here, an agent is presented with two objects, one after the other, and its task is to catch the second object if it is smaller than the first and to avoid it otherwise. Again, this task can be described in similar terms: each of the two objects is either *first* or *second* and either *larger* or *smaller*, and the appropriate behavioral response requires that these two sets of features be aligned properly.

## 6.2 Modeling Relational Categorization

Models of relational categorization have traditionally taken its role-filler *description* as a reflection of the *solution* to the categorization problem, by assuming explicit representations for each of the features and focusing on mechanisms for binding them. For instance, in the numerous connectionist and hybrid symbolic-connectionist models of relational categorization that have been developed, features are represented explicitly either as symbols or as distributed patterns of activity over sets of neural units [67, 101–103, 116, 229]. A natural consequence of adopting this approach is the so-called binding problem, which concerns how features can be associated with each other in such a way that the two possible bindings are distinguished. Diverse solutions to the binding problem have been proposed, as reviewed in [67]. One common solution is to dynamically mark separate role and filler elements as belonging together [1, 102, 104, 207]. That is, in addition to an activation, elements have another associated value which, when it matches that value for another element, represents a binding between them. An alternative solution is to incorporate role information directly into the encoding of a set of input values. For example, a feature vector and a special "role" vector may be combined—e.g., using convolution [178] or the tensor product [211]—to form the inputs to a connectionist network [89]. In this way, binding is implemented through an explicit combination mechanism. A third solution is to allocate different parts of the relational system—e.g., separate banks of units in a connectionist network—to represent different roles [88].

A key point about all of these solutions is that they rely on a conception of relational categorization as an abstract computational task, in which information is represented as a set of abstract features and is processed in a stepwise computational procedure. These assumptions allow modelers to make ad hoc decisions in designing their representations or network architectures to deal with the binding problem. In contrast, our approach to relational categorization makes no such assumptions. Instead of studying relational categorization in terms of abstract representations for the features belonging to a high-level description of the problem (*first*, *second*, *larger*, *smaller*) and a way to bind these representations together, we focus on grounded relational behavior in embodied agents. A crucial advantage of this approach is that it significantly broadens the playing field for possible relational mechanisms. In particular, the analysis presented below will demonstrate how relational categorization may be carried out without any obvious binding mechanisms, thus avoiding consideration of the binding problem altogether.

Several other models are noteworthy for eschewing the binding view of relational categorization and, like the models studied here, demonstrating the ability of dynamical neural circuits to perform relational tasks. For example, in one study a simple recurrent network was trained to recognize string sequences of the form $a^n b^n$, and thus to identify a *same count* relationship between sequences of inputs [188]. However, the task performed by this model, as with those performed by binding models, is disembodied and computational in nature, whereas the work here is concerned with relational behavior in situated and embodied agents. In another study, a large-scale spiking network model was developed to capture experimental findings from a relational task performed by macaque monkeys [26,155,189]. In this case, though, the relational mechanism was hand-designed, while in the work presented here we employ evolutionary techniques, thereby attempting to minimize a priori assumptions about how the relational mechanisms should work.

The modeling approach used here has been applied to a wide variety of behaviors, including chemotaxis and walking [19], navigation [242], object categorization [16,17,244], short-term memory [210], associative learning [177], decision making [237], selective attention [77,210,245], agency detection [51], and communication [184,253]. The central idea of this approach is to use evolutionary algorithms to evolve "nervous systems" for embodied and situated model agents as a way to develop the conceptual foundations and theoretical tools necessary for understanding brain-body-environment systems [20,33,50,94,162]. Since animals were evolved, not designed, they were selected for their overall behavioral efficacy rather than their understandability. Thus, by mimicking the process by which biological brain-body-environment systems were produced, we can likewise evolve model agents that can exploit the freedom to partition solutions across brain-body-environment boundaries in ways that will not necessarily align with our preconceptions about how such systems should work. However, our evolved model brain-body-environment systems have the significant advantage that we have complete access to and control over their nervous systems, bodies, and environments, allowing us to undertake detailed analyses of their operation. In addition, many analysis techniques, including those of information dynamics, are quite data-intensive, and it is better to understand the strengths and limitations of these techniques in a context where data is plentiful before turning to the approximations necessary for working with empirical data.

## 6.3 Methods

The model agent and environment used in this study are essentially the same as those used in previous work on object categorization [16,17]. The agent has a circular body with a diameter of 30[1], and an array of 7 distance sensors equally spaced over an angle of $\frac{\pi}{6}$ radians on the agent's top side (Figure 6.1(a)). Each distance sensor has a maximum length of 220. Distance sensors take on values that are inversely proportional to the distance at which their corresponding rays intersect objects in the environment, with a minimum value of zero when rays are at their maximum length and a maximum value of 10 when rays are at zero length (i.e., when an object is immediately above the agent). The agent is positioned along the bottom edge of a planar environment and is able to move horizontally with a maximum velocity of 5 in either direction. The environment extends indefinitely in both horizontal directions, so that the agent's motion is unimpeded by environmental boundaries. Circular objects fall towards the agent from above with a constant vertical velocity of -3. As will be elaborated later, the agent's task is to "catch" or "avoid" various of these objects, where catching or avoidance is determined by the horizontal separation between the agent and object when the object completes its fall.

The agent's behavior is controlled by a continuous-time recurrent neural network [15, 18], which is a particular instance of the general class of additive neural network models that have been extensively studied in neuroscience [86,95,99,100,170]. Each unit in the network is governed by the state equation

$$\tau_i \dot{s}_i = -s_i + \sum_{j=1}^{N} w_{ji}\sigma(s_j + \theta_j) + I_i \qquad i = 1, \ldots, N \tag{6.1}$$

[1] Note that the units of measurement are essentially arbitrary. For concreteness, one can assume that distances are in centimeters, time is in seconds, and velocities are in cm/sec.

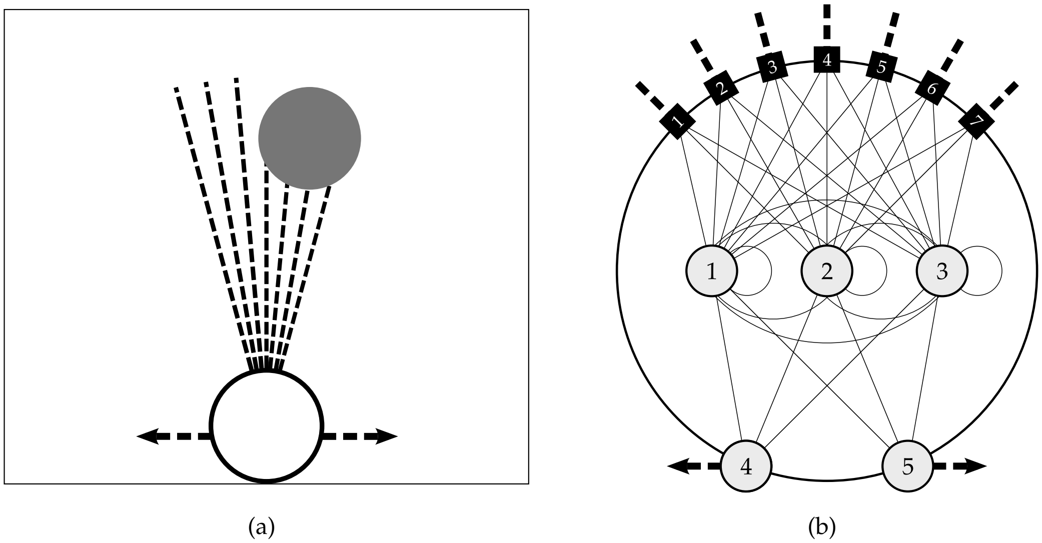


Figure 6.1: The agent and environment. (a) The agent moves horizontally while circles fall towards it from above. The agent's sensory apparatus consists of an array of seven distance sensors. (b) The distance sensors fully project to a layer of fully interconnected interneurons, which in turn fully project to the two motor neurons.

where $s_i$ is the state of neuron $i$, $\tau_i$ is its time constant, $w_{ji}$ is the strength of the connection from the $j^{th}$ to the $i^{th}$ neuron, $\theta_j$ is the bias for neuron $j$, $\sigma(x) = \frac{1}{1+e^{-x}}$ is a logistic activation function, and $I_i$ represents an external input to neuron $i$. The output of neuron $i$ is given by $o_i = \sigma(s_i + \theta_i)$. The standard neurobiological interpretation of this model is that $s_i$ represents the mean membrane potential for neuron $i$, $\sigma(\cdot)$ represents its mean firing rate, $\tau_i$ represents its membrane time constant, $\theta_i$ represents its threshold, the weights $w_{ji,j\neq i}$ represent synaptic connections from neuron $j$ to neuron $i$, and $w_{ii}$ represents a simple active conductance. Alternatively, the model can also be interpreted as representing nonspiking neurons [54] or subcellular signalling pathways [46,212], or can be viewed simply as a convenient and extremely general basis of dynamics[2]. The neural architecture for an agent is depicted in Figure 6.1(b). The seven sensors (black boxes 1-7) are fully connected to a layer

[2] Indeed, continuous-time recurrent neural networks have been shown to be universal approximators of smooth dynamics [31,65,118], so that in principle they can reproduce the behavior of any smooth dynamical system to an arbitrary degree of accuracy.

of three interneurons (circles 1-3), which are fully interconnected and project fully to two motor neurons (circles 4-5). The agent's horizontal velocity is proportional to the difference between the outputs of the two motor neurons. Thus, the equations for the complete model take the form

$$\begin{aligned}
\dot{y} &= -3 \\
\tau_i \dot{s}_i &= -s_i + \sum_{j=1}^{3} w_{ji}\sigma(s_j + \theta_j) + \sum_{j=1}^{7} I_j(x, y, d) \quad i = 1, 2, 3 \\
\tau_i \dot{s}_i &= -s_i + \sum_{j=1}^{3} w_{ji}\sigma(s_j + \theta_j) \quad i = 4, 5 \\
\dot{x} &= 5(\sigma(s_4 + \theta_4) - \sigma(s_5 + \theta_5))
\end{aligned} \tag{6.2}$$

where $x$ is the horizontal position of the object relative to the agent's midline, $y$ is the vertical position of the object relative to the agent, and $I_j(x, y, d)$ is the sensory input from the $j^{th}$ distance sensor due to a circular object with diameter $d$ at location $(x, y)$ in agent-centered coordinates. Models were simulated using the forward Euler method with an integration step size of 0.1.

The evolutionary algorithm used in this study is essentially the same as that described in [16]. The algorithm acts on a population of real-valued vectors, where each vector represents the neural parameters for an individual agent, including its time constants, biases, and connection weights (from sensors to neurons and between neurons). The components for each vector are mapped to neural parameters using linear maps from ±1 to the following ranges for each parameter: time constants $\in [1, 30]$, biases $\in [-16, 16]$, and connection weights $\in [-16, 16]$. Additionally, time constants are clipped to be greater than or equal to 1, so that they never become too small relative to the integration step size. At the outset, a population is initialized by assigning each component of every vector a random value uniformly distributed over the range ±1. Then, the fitness of each individual is evaluated using a method that will be described momentarily, and individuals are selected for reproduction using fitness proportionate selection with linear fitness scaling and a fitness scaling multiple of $1.01$ [11, 76]. For each selected individual, a "child" is produced via mutation by adding to it a random vector whose direction is uniformly distributed on the $M$-dimensional hypersphere ( [126], p. 130) and whose magnitude is normally distributed with a mean of 0 and a variance of 4. Finally, for each slot in the next generation, the child is chosen if its performance is greater than or equal to that of its parent, and otherwise the parent is copied.

Agents are evolved for the ability to make discriminations based on the relational category *smaller*. This ability is assessed in a task wherein agents are shown pairs of circular objects, presented one after the other, and the objective is to catch the second circle in each pair if it is smaller than the first, and to avoid it otherwise[3]. Crucially, the same circle may be either *smaller* or *larger* depending on the other, so the agent must attend to the size relation between the two circles. An agent's performance in this task is determined based on its behavior in a number of evaluation trials, where each trial proceeds as follows. First, the agent's neural states are initialized to 0. Then, a circular object with diameter $\in [20, 50]$ begins falling from the top of the environment with a horizontal offset of 0 and a vertical offset of 220 relative to the topmost point on the agent's body. Thus, the object begins its fall along the agent's midline and just beyond its field of view. The object falls until it reaches a vertical position coinciding with the top of the agent and is then removed from the environment. Then, a second circular object begins falling from the same initial position as the first object. The second object also has diameter $\in [20, 50]$, but with this value differing from that of the first object by at least 5 so that the size difference between the two objects will be perceptible to the agent given the coarse spatial resolution of its sensors. The second circle falls until it reaches the agent, and then the final horizontal separation between the agent and the second object is recorded. This final separation constitutes the

[3] Interestingly, a very similar task was used in an earlier study on imprinting behavior [107].

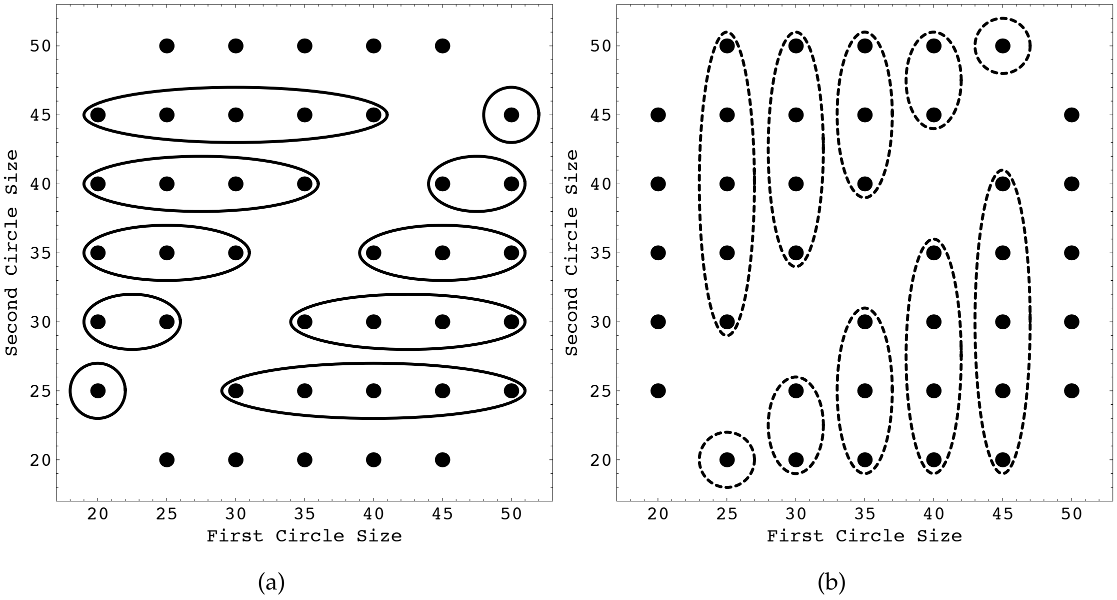
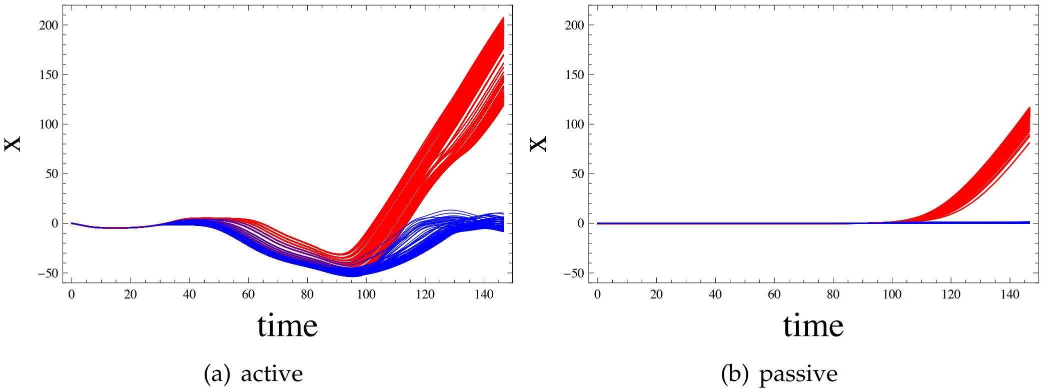


Figure 6.2: Procedure for fitness evaluation. Black dots indicate the 40 evaluation trials. Scores on these trials are combined to form two aggregate measures, represented in (a) and (b), respectively. These two measures are then combined to determine overall fitness. See the main text for details.

agent's catch/avoid response, with a separation of 0 corresponding to a perfect catch and a separation of *MaxDistance* (= 75) or greater corresponding to perfect avoidance. The agent's score on a trial is $1 - d$ if the second circle is smaller and $d$ if the second circle is larger, where $d$ is the final separation clipped to *MaxDistance* and normalized to $[0, 1]$.

The fitness of each agent is calculated from a total of 40 evaluation trials, where each trial consists of a different combination of first and second object sizes (black dots in Figure 6.2). Initially, we tried simply averaging the scores on all trials in order to determine the overall fitness for an agent. However, assigning fitness in this way resulted in agents adopting one of two suboptimal solutions: agents would catch or avoid based on the size of either the first or second object alone (e.g., always catching if the first object size was greater than 35, and avoiding otherwise), and thus would ignore the size relation between the two objects. To compensate for this, we used a slightly more sophisticated procedure



(a) active

(b) passive

Figure 6.3: Active and passive behavioral strategies. Sample trajectories of the agent's x-position for the (a) active and (b) passive agents. Blue trajectories correspond to trials where the second object is smaller than the first (catch trials), while red trajectories correspond to trials where the second object is larger (avoid trials).

for weighting and averaging the trial scores. First, an aggregate measure called *rowAgg* was calculated by averaging the scores in each solid oval in Figure 6.2(a), resulting in ten numbers, and then taking the average of those ten numbers. Thus, for each second object that appears in at least one trial where it is smaller and one trial where it is larger, *rowAgg* assigns equal weight to smaller and larger trials. In this way, *rowAgg* penalizes strategies based on observing the size of the second object only. Next, an analogous procedure was carried out for the dashed ovals in Figure 6.2(b) to calculate a second aggregate measure *colAgg*. Thus, *colAgg* penalizes strategies based on observing the size of the first object only. Overall fitness was then computed as the minimum of *rowAgg* and *colAgg*.

## 6.4 Results

We performed ten evolutionary runs each for agents having five, four, and three interneurons, respectively. In each run, a population of 200 individuals was evolved for 1,000 generations. The best agent in each run had a fitness of at least 90% on 10,000 random size combinations. A further attempt to evolve agents with two interneurons was unsuccessful, suggesting that three interneurons may be the minimum necessary for high performance on this task.

The behavioral strategies of the agents with three interneurons fall into two categories. One group of agents employ an active strategy, moving back and forth repeatedly to scan circles as they fall (Figure 6.3(a)). The other group use a passive strategy, remaining mostly still as the first circle falls, and then either remaining in place to catch the second object or veering to one side and avoiding it (Figure 6.3(b)). For the remainder of this chapter, we will focus exclusively on the passive strategy, though in Chapter 7 we will also analyze some aspects of the active strategy. This decision to concentrate our preliminary analysis on the passive strategy was based on three considerations: (1) the agent with the best performance uses it; (2) the majority of agents (6 out of 10) use it; (3) the relational mechanism underlying it is more straightforward. In particular, we will focus on the best evolved agent as a characteristic example of the passive strategy. The best agent had a fitness of 99.83% over 10,000 random size combinations. A quick glance at the agent's performance over the range of object sizes verifies the desired pattern for relational categorization (Figure 6.4), with the agent's behavior divided into distinct catch and avoid regions coinciding with smaller and larger second objects, respectively.

## 6.5 The Dynamics of Relational Categorization

To understand what produces the agent's relational behavior, it is necessary to examine the agent's neural dynamics. In particular, two features of the dynamics turn out to be essential. The first feature has to do with how the agent stores the size of the first object. The agent must store this size in order to respond differently to the second object when it is smaller or larger. This means that some aspect of the agent's state—its neural activations and horizontal position—must correspond to the size of the first object at the

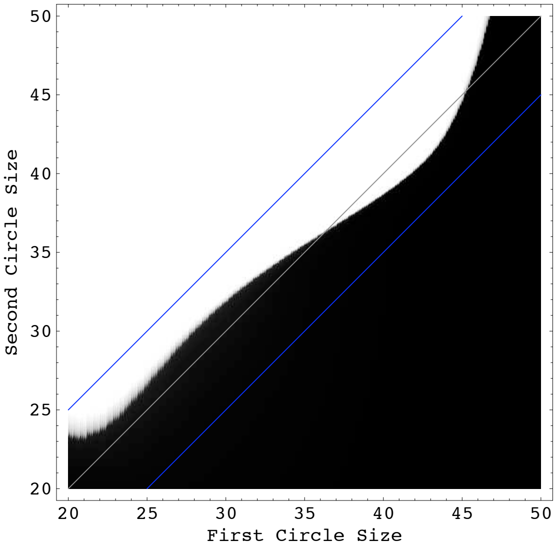


Figure 6.4: Performance of the best evolved agent. A density plot of the agent's final horizontal separation from the second object, with black indicating a perfect catch and white indicating perfect avoidance. The gray line indicates where the two object sizes are equal. The blue lines indicate where the size difference between the two objects equals the minimum size difference used in evaluation trials.

trial midpoint (i.e., when the first object is removed and the second object begins its fall). In general, this correspondence could be quite complex and nonlinear, incorporating any subset of the agent's state variables. However, in the agent under consideration the size information is stored directly; namely, the output of one interneuron, Neuron 3 (N3), stores the size. The output of N3 linearly correlates with the size of the first object at the trial midpoint (Figure 6.5), whereas all other state variables take on essentially constant values. We verified this property of N3 by setting the other state variables to suitable fixed values at the trial midpoint and confirming that performance changed negligibly as a result.

The second feature of the dynamics has to do with what initiates the agent's catch/avoid response. The crucial observation here is that another interneuron, N1, either switches off (i.e., its activation decays to zero) or remains on during the fall of the second object (Figure 6.6). Furthermore, whether or not N1 switches off determines the agent's catch/avoid

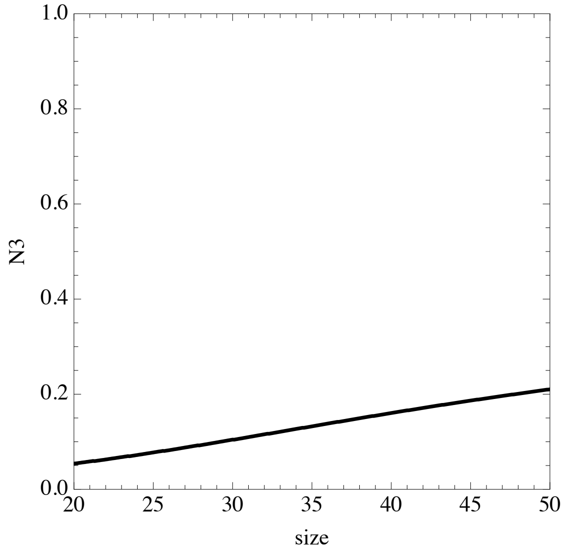


Figure 6.5: Output of N3 at the trial midpoint. The output of N3 is plotted as a function of first object size. The output of N3 stores the size by linearly correlating with it.

response. If N1 switches off, the agent catches the second object, while if N1 remains on the agent avoids it. Thus, we can account for the agent's relational behavior by understanding what factors determine whether or not N1 switches off. Consequently, we next need to explore how these two features of the dynamics—the stored size in N3 and the switching behavior of N1—interface with one another. Specifically, how does the activation of N3 combine with the influence of the second object to trigger the switching behavior of N1, and as a result the agent's catch/avoid response?

To probe this question, we first examine the effect of N3 on N1. This effect is determined primarily by the connection from N3 to N1, which is negative (i.e., inhibitory). Thus, N3 tends to cause N1 to switch off, and greater activations of N3 cause N1 to switch off more quickly (Figure 6.7). Since N3 stores the size of the first object, this means that larger first object sizes tend to produce faster decays in the output of N1. However, despite this tendency of N3 to switch N1 off, we know that in some cases N1 remains on, and that whether or not N1 switches off determines the agent's catch/avoid response. So, what

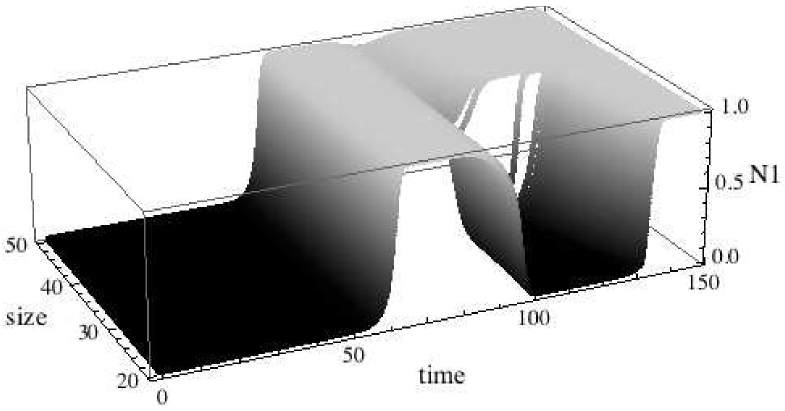
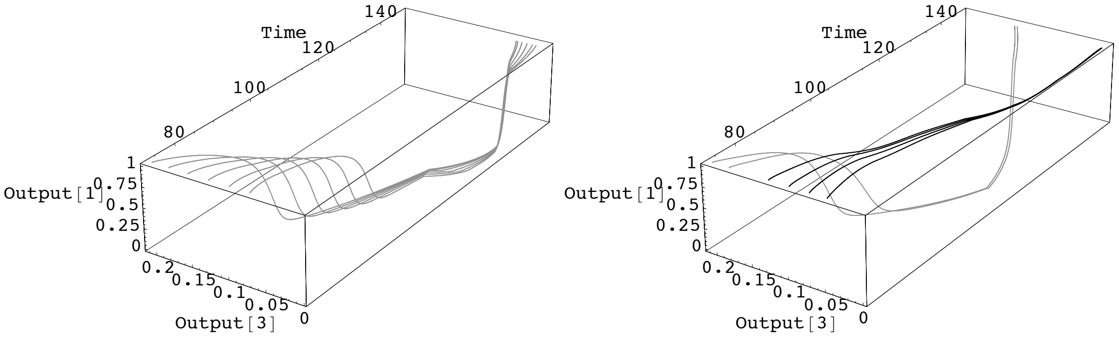


Figure 6.6: Switching behavior of N1. Trajectories of N1 corresponding to a first object size of 35 for a range of second object sizes. Whether or not N1 switches off determines the agent's catch/avoid response.

determines whether or not N1 switches off? To answer this question, we must explore the underlying equilibrium structure of the neural dynamics.

The neural circuit is a nonautonomous dynamical system, since it receives time-varying inputs from the distance sensors. A common strategy to analyze a nonautonomous system is to examine its autonomous dynamics when the inputs are held constant, and then to examine how this dynamics varies for different sets of inputs. The nonautonomous dynamics can then be approximated by considering the sequence of autonomous systems corresponding to the particular inputs that the nonautonomous system receives. In our case, three factors determine the inputs to the agent's neural circuit: (1) the size of the object; (2) the vertical offset of the object; (3) the horizontal offset of the object. However, since the agent remains still for most of each trial (i.e., it uses the passive strategy), we can simplify our analysis by ignoring changes in the input due to (3). Thus, we can explore the dynamical structure of the neural circuit by considering it as a function of two variables: the size of the object and its vertical separation from the agent.

It turns out that the only limit sets exhibited by the neural circuit are equilibrium points,

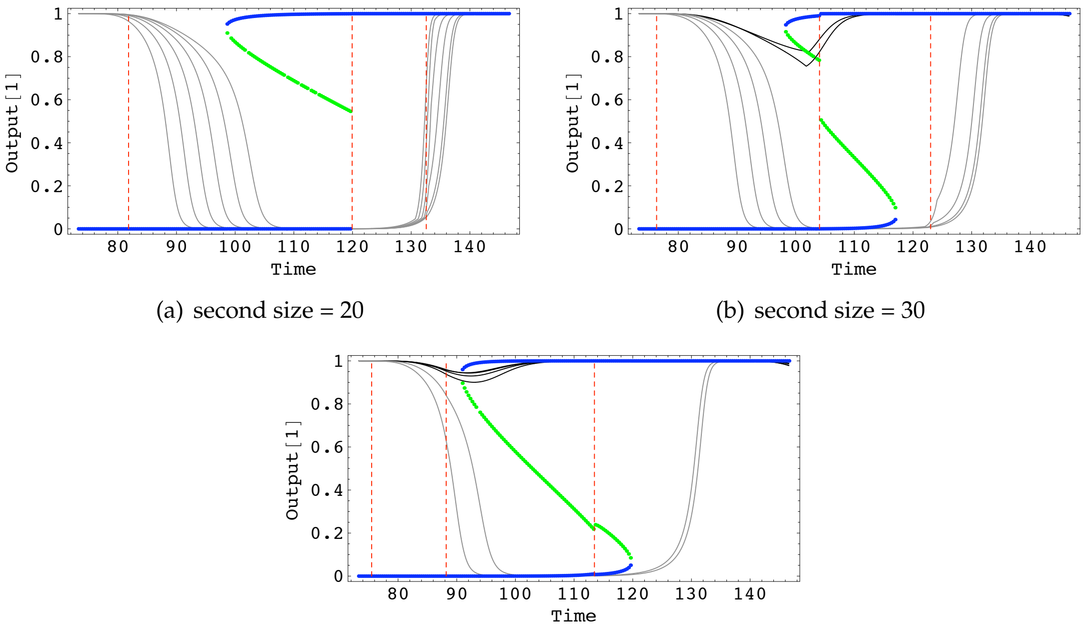


(a) second size = 20

(b) second size = 40

Figure 6.7: The impact of N3 on N1. Trajectories of N1 and N3 over the second half of a trial for several first object sizes. Gray trajectories are catch trials and black trajectories are avoid trials. Higher activations of N3 cause N1 to switch off more quickly.

whose positions and stabilities we can calculate [222]. The circuit has a single stable equilibrium point when the object is very far or very close to the agent, but is bistable when the object is at an intermediate distance, with saddle-node bifurcations separating these unistable and bistable regions. Thus, at intermediate distances, the circuit has two stable equilibrium points, which serve as attractors for system trajectories, and a single saddle point, which separates the two basins of attraction. How does this equilibrium structure explain the switching behavior of N1? To answer this question, we can superimpose representative trajectories of N1 over a plot of the system's equilibrium points (Figure 6.8), showing how the trajectories of the system are shaped by the underlying organization of its dynamics. Note that the bistable region separates the trajectories of N1 into those that switch off and those that remain on, corresponding to catch and avoid trials, respectively. The bistable region is produced by a bifurcation that occurs when the second object reaches a certain vertical offset. The timing of this bifurcation correlates with the size of the second object (occurring earlier for larger second objects), so that trajectories of N1 are appropriately partitioned into catch and avoid regions.



(a) second size = 20

(b) second size = 30

(c) second size = 40

Figure 6.8: The influence of bistability on N1. Trajectories of N1 for a range of first object sizes (gray and black lines) along with the equilibrium points (EPs) of the neural circuit. Blue EPs are stable and green EPs are saddles. The bistable region separates the trajectories into two basins of attraction. The dashed red lines show where the agent's ray sensors are first broken. Other abrupt changes in the underlying dynamics are due to bifurcations. Note that there is a minor inconsistency between the trajectories and EPs due to the fact that horizontal offsets between the agent and object were ignored.

With these results, we can now assemble a dynamical account of the agent's relational mechanism as follows. During the first half of a trial, the output of N3 comes to correlate with the size of the first object, while N1 switches to fully on. Then, during the second half of a trial, N1 begins to switch off. Due to the inhibitory effect that N3 has on N1, N1 switches off more quickly for larger first objects. As the second object falls, a bifurcation occurs whose timing depends on the size of the second object. If the activation of N1 is low enough by the time the bifurcation occurs, N1 switches off completely and the second circle is caught. Otherwise, N1 returns to fully on and the second circle is avoided. The inhibitory effect of N3 on N1 and the timing of the bifurcation are coordinated such that smaller second objects are caught and larger second objects are avoided.

## 6.6 Discussion

One significant feature of the analysis presented above is that it demonstrates the crucial role that time can play in a dynamic neural mechanism. For example, the activation of N3, which stores the size of the first object, affects the subsequent network dynamics by causing N1 to switch off more quickly for larger first objects. Also, the size of the second object determines the timing of when a bifurcation occurs in the system, dividing the trajectories into catch and avoid responses. Finally, the relational mechanism as a whole relies on the coordinated timing of N1 switching off and the occurrence of the bifurcation. The central importance of time runs as a common thread throughout these examples.

The analysis also shows how information can be "represented" and "processed" in a variety of ways in a dynamic neural circuit. One piece of information, the size of the first object, is stored in the activation level of N3. This encoding of a perceptual feature by a neural activation fits with the standard approach to representation in connectionist networks. However, if we try to identify what serves to represent the size of the second object, the only obvious candidate is the bifurcation that the system undergoes. This correspondence between a stimulus feature and the onset of a bifurcation is an entirely different way for information to bear on the network dynamics. Nevertheless, the analysis makes clear how these different features are integrated seamlessly to produce relational categorization.

As we will see in the next chapter, other agents use their position in the environment to store information about object sizes. Thus, in this case the position of the agent, as well as its neural dynamics, plays a crucial role in the relational mechanism. As we will discuss, this kind of strategy fits nicely with one popular idea about the importance of situatedness and embodiment for cognition, which is that they allow a cognitive agent to offload information to its environment. One strength of dynamical approaches is that neural, bodily, and environmental variables are all represented in the same dynamical language, so that cognitive processes may naturally spread across the entire brain-body-environment system. Similarly, our information dynamics approach described in the next chapter applies just as easily to bodily and environmental variables as it does to neural ones, allowing us to track the flow of information for embodied and extended solutions to cognitive tasks.

# 7

# Information Dynamics of Embodied Relational Categorization

In this chapter, we apply our information dynamics framework to analyze the models of relational categorization described in the previous chapter. First, in Section 7.1, we provide the details of our approach, describing how the tools introduced in Chapter 5 are applied to our model agents. Then, in Section 7.2, we analyze several features of the passive agent that was the focus of Chapter 6, including how the agent extracts information about the sizes of objects and how the agent integrates information about the sizes of different objects. In Section 7.3, we then turn to an analysis of an agent using an active behavioral strategy for relational categorization. The analysis of this agent illustrates several of the unique ways that embodiment can influence information dynamics, including the ability of embodied agents to actively elicit and structure sensory information, and their ability to offload information to their bodies and environments. Finally, in Section 7.4, we conclude with some general discussion and directions for future work.

## 7.1 Information Dynamics of Embodied Agents

The central idea of our information dynamics approach is to use the tools developed in Section 5.2 to explore how information about particular stimulus features flows through a brain-body-environment system. For example, in our analysis of the relational agents,

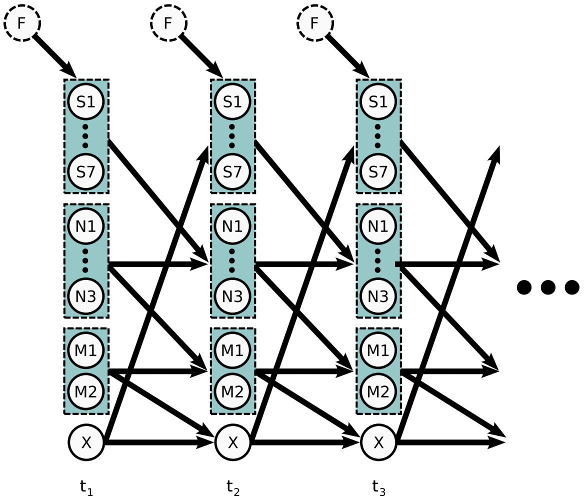


Figure 7.1: Brain-body-environment system as a dependency graph. Each component of the brain-body system (sensors S1-S7, neurons N1-N3, motors M1-M2, and body position X) is represented by a time-indexed sequence of nodes, with a directed arrow from one node to another representing a causal influence of the former onto the latter. The influence of environmental features is represented by the nodes labeled $F$.

the stimulus features that we will consider are the sizes of the first and second objects presented to an agent. Traditionally, information-theoretic analyses have focused on static or time-averaged measures of informational structure. In contrast, our approach is to unroll these static measures across time to explore how informational structure evolves over the course of behavior. In addition, by further unrolling across values of the stimulus feature, we are able to trace how information about particular stimuli flows through the system.

To conceptualize our approach, it is helpful to visualize a brain-body-environment system as a dependency graph (Figure 7.1). In such a graph, each component of the brain-body system is represented by a time-indexed sequence of nodes (circles in Figure 7.1), with a directed arrow from one node to another representing a causal influence of the former onto the latter. For example, the dependency graph in Figure 7.1, representing the network of dependencies for one of our relational agents, includes nodes for all sensory (S1-S7), neural (N1-N3), motor (M1-M2), and bodily (X; the agent's horizontal position) components of the agent's brain and body. Additionally, the influence of environmental variables, such as the size of a falling object, is represented by the nodes labeled $F$, with the arrows from these nodes to those of the sensors representing the influence of environmental features on the sensory input that the agent receives. Also note the cyclic progression of arrows from $S \rightarrow N \rightarrow M \rightarrow X \rightarrow S$, which represents the feedback relationship between sensory inputs and the agent's motion. In the context of such a diagram, the aim of our approach can be interpreted graphically as that of tracking the flow of information about $F$ along various paths through the dependency graph. For example, the question of how information flows through the neural components can be answered by tracking the information about $F$ along the path $N_t \rightarrow N_{t+1} \rightarrow N_{t+2} \rightarrow \cdots$. Similarly, to answer the question of how the agent's motion shapes its sensory input requires an understanding of the path $M \rightarrow X \rightarrow S$.

Concretely, the first step in our analysis is to evaluate an agent's behavior for a uniformly distributed sample of the stimulus feature, recording the trajectories of all neural and bodily state variables for each stimulus presentation. From the values taken on by each state variable at each moment in time and the corresponding stimuli that produced them, we then estimate a joint probability distribution over values of the state variable and the stimulus feature. In other words, the stimulus feature is treated as a random variable $F$, and each state variable of the agent system (its sensors, neurons, and body position) is treated as a stochastic process—a time-indexed sequence of random variables—and we estimate joint distributions for $F$ paired with each time-indexed random variable. To estimate these distributions, we use average shifted histograms [199], a simple form of kernel density estimation, because they provide a beneficial trade-off between computational and statistical efficiency, though other density estimation techniques could certainly be used [200, 208]. In particular, for our analysis of the relational agents, each agent was evaluated on object sizes in the range [20, 50] (the same used during evolution) sampled at an interval of 0.01. Probability distributions were then estimated using average shifted histograms with 100 bins and 8 shifts along each dimension, though all reported results were qualitatively robust over a broad range of discretizations (20 to 200 bins).

Our analysis then proceeds by calculating informational measures on the basis of these estimated probability distributions. First, we calculate the mutual information $I(F; V_t)$ for each state variable $V$ and time index $t$, thereby generating a flow of information for each state variable of the system. To better understand the dynamic properties of these flows, we next apply our measure of information gain (Equation (5.4)) to the flow for each state variable. Then, to determine the sources of information gain, we quantify the information transfer (Equation (5.5)) originating from each of the causal influences onto a state variable. Finally, to explore the structure of information flow at a finer level of detail, we use the concept of specific information (Section 2.4) to expand our analysis. Specifically, we unroll our measures of information flow $I(F; V_t)$ to consider the specific information $I(F = f; V_t)$ that each state variable $V$ provides about each particular stimulus $f$ at each time $t$. Similarly, applying specific information to expand our measures of information gain and transfer, we explore how state variables acquire and exchange information about particular stimuli as a function of time.

## 7.2 Analysis of Passive Relational Categorization

As a first demonstration of our information dynamics approach, we next apply it to explore two questions about the passive relational agent discussed at length in Chapter 6. First, we examine how the agent extracts information about the size of the first object. Then, we explore how the agent integrates size information from the first and second objects in order to make its relational discrimination.

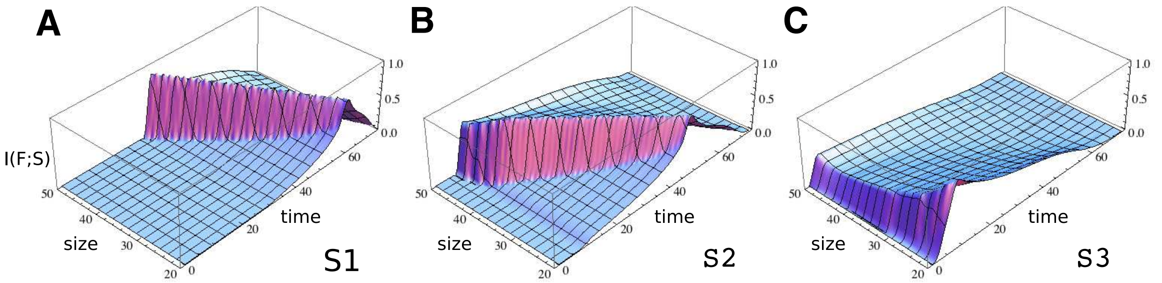


Figure 7.2: The flow of first object size information for sensors S1-S3. All plots throughout this chapter are normalized to [0,1] by dividing by the maximum possible value for specific information.

We begin by examining what information about first object size (henceforth $F$) is available to the agent. In general, this amounts to examining the flow of information about $F$ in each of the agent's seven sensors. However, since the agent uses a passive strategy and objects are always presented at the agent's midline, information from the sensors is bilaterally symmetric and thus we can simplify our analysis by considering only the sensors on one side. Plots of the information flow for sensors S1-S3 are shown in Figure 7.2, where specific information is plotted as a function of object size and time. The flow of information for each sensor exhibits the same prominent feature: a high "ridge" of information that begins first for large sizes and travels across to successively smaller sizes. However, the time at which this ridge forms varies for the three sensors, beginning first for the innermost sensor (S3) and later for sensors further from the agent's midline.

What do these plots tell us about the information available to the agent? Recall that a high value for the specific information indicates that a particular object size is unsurprising given the state of the sensor at a certain point in time. This means that the sensor tends to take on distinct values for that object size, serving to distinguish that object size from differently sized objects. Thus, the high ridge of information informs us that the state of each sensor first distinguishes large objects, then objects successively smaller in size. Furthermore, by examining the path of the object with respect to the agent, the origin of these ridges becomes immediately clear: the peak value in information for each object size occurs at precisely the time when objects of that size first intersect the corresponding ray sensor. Since larger objects intersect earlier, the sensors first provide information about these sizes. Thus, the primary information available to the agent stems from the timing of when differently sized objects intersect each ray sensor, with this timing information varying for each of the three sensors.

The next step is to inspect how size information flows through other components of the brain-body-environment system. In general, this includes both the interneurons and the position of the agent's body; however, since the agent uses a passive strategy, we can simplify our analysis by considering only neural state variables. Each of the three interneurons shows a markedly different pattern of information flow (Figure 7.3). First, consider the information contained by each neuron at time $73.5$, which corresponds to the time when the first object is removed. Previously, in Chapter 6, we noted the primary role played by N3 in storing information about $F$, a fact that is also evident from the plots of information flow. When the object is removed, N3 contains a high amount of information about all object sizes, while N1 and N2 contain no information. However, crucially, while our previous observation about N3 was based on a single snapshot of the agent's state, the plots in Figure 7.3 depict the entire temporally-extended process that leads to this eventual state. Thus, by examining information flow, we can move beyond considering simply how information ends up stored, and instead explore the dynamic process that results in this storage. In particular, the plots in Figure 7.3 make clear that, although size information ends up concentrated in N3, both N1 and N2 also contain size information at earlier times in the trial. N1 contains a high amount of information about small and large sizes late in the trial, while N2 contains information first about small and then about large sizes. Thus, contrary to the picture of N3 as the sole bearer of size information, these observations suggest that N1 and N2 may also play an important informational role, an idea that we return

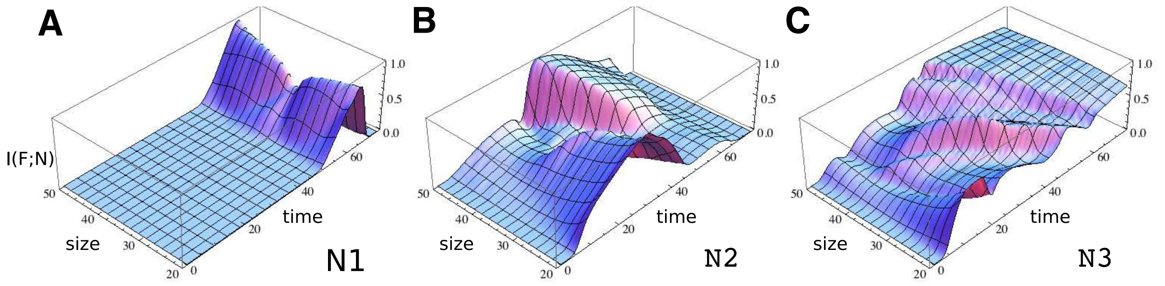


Figure 7.3: Flow of first object size information for neurons N1-N3 of the passive agent.

to momentarily.

Nonetheless, the most striking feature of these information flow plots is the gradual build-up of size information in N3 (Figure 7.3C). In contrast with N1 and N2, which both gain and subsequently lose information, N3 continually accumulates information through a succession of information "waves". To better understand the source of these waves, we next examine the dynamics of information gain for N3 (Figure 7.4). Comparing Figures 7.3C and 7.4 reveals that the waves of information flow translate to ridges of information gain. In particular, the plot of information gain shows a series of four prominent ridges, each traveling across from large to small sizes, as well as some other secondary ridge-like features. Importantly, the prominent ridges of information gain closely align with the ridges of information flow produced by the agent's sensors (Figure 7.2). Recall that the latter are produced by objects intersecting the ray sensors at different times depending on their size. Thus, together these observations suggest a hypothesis for how the agent

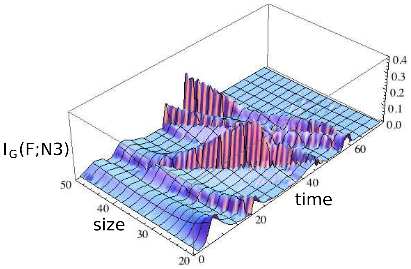


Figure 7.4: Information gain for N3.

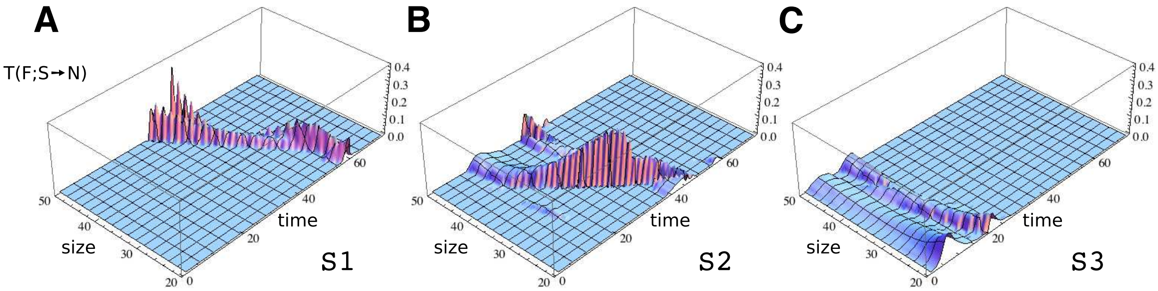


Figure 7.5: Transfer of first object size information for sensors S1-S3 to neuron N3.

extracts size information: N3 primarily integrates information available from the sensors, stemming from the timing of when each sensor is broken by differently sized objects.

To explore this hypothesis, we next calculate the transfer of information about first object size from each of the three sensors to N3 (Figure 7.5). The results of these calculations provide strong evidence in support of our hypothesis, confirming the relationships between sensor information flow and information gain in N3 that can be gleaned from a visual comparison of Figures 7.2 and 7.4. In particular, each of the four prominent ridges of information gain can be accounted for by one of the agent's sensors, with the two ridges occurring earliest in time linked to S3, and the ridges occurring third and fourth associated with sensors S2 and S1, respectively. However, one must be cautious when interpreting these results, as information that appears to be transferred by one component may also be provided redundantly by another. In this sense, our measure of extrinsic information transfer, like transfer entropy, is actually a measure of *apparent* transfer, as its redundant portion is ambiguous with regards to the underlying causal structure (recall the discussion in Section 5.1). Indeed, calculating the information transfer from N2 to N3 (Figure 7.6), we find that this sort of causal ambiguity is exactly the case. Specifically, the information transfer from N2 includes portions of three of the four prominent ridges of information gain, indicating that this information may be transferred either from N2 or from the sensors.

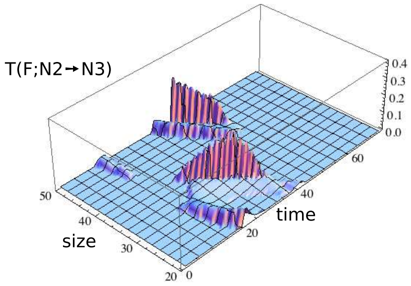


Figure 7.6: Transfer of first object size information for neuron N2 to neuron N3.

Unfortunately, this ambiguity cannot be resolved using purely informational measures—those based only on observing, rather than perturbing, a system—as the determination of causal relationships requires some method for intervening in the system.

Consequently, to test our hypothesis regarding how information accumulates in N3, we next explored the impact of removing information from particular components using an interventional method called *informational lesioning* [115]. The basic idea of informational lesioning is to systematically diminish a component's ability to convey information while simultaneously minimizing its deviation from normal behavior. In this way, the informational contribution of a component can be quantified independently from other functional roles that the component might play.

We first applied this idea to remove the information provided by each sensor. For each of the sensors, normal behavior is essentially the same: it remains off for a certain period before intersecting the object, then begins to increase monotonically, with the timing of this transition dependent on the size of the object. Additionally, the rate at which sensor values increase, i.e., the curvature of the sensory trajectory, also varies with object size. Thus, size information actually comes from two features of the sensory input: the timing of the off/on transition, and the curvature of the subsequent trajectory. In order to isolate timing information, which our analysis predicts to be the most salient, we first removed information due to varying curvatures. This was done by replacing the curved portion of sensor trajectories with a single best-fit linear approximation, formed separately for each sensor, so that trajectories for a given sensor differed only in the timing of their off/on transition. Performing this manipulation simultaneously on all sensors had essentially no impact on performance (99.67%), thus confirming our prediction that timing information is primarily what the agent uses to extract object size.

Next, to test the prediction that timing information from different sensors is integrated, we independently removed the timing information from each symmetric pair of sensors by setting their values according to the mean off/on transition time. Under this manipulation, performance dropped to 71.84% for S1/S7, 92.53% for S2/S6, and 96.93% for S3/S5. Thus, timing information clearly is integrated from the different sensor pairs, though with varying contributions made by each. Interestingly, the relative contribution from each pair correlates with the magnitude of information gain that each produces for N3 (height of ridges in Figure 7.4). For example, the greatest impact on performance results from lesioning S1/S7, which also produce the largest information gain for N3. Thus, not only does information flow analysis yield a correct qualitative prediction for how information is extracted, but also points to some quantitative features of this process.

However, as alluded to earlier, this explanation does not provide the full story. In particular, the other two interneurons also contain size information at different times, and thus may also contribute to the information stored in N3. Moreover, the results conveyed by Figure 7.6 indicate that N2 may transfer information directly to N3, with N2 potentially accounting for several features of N3's information gain. To explore these possibilities, we performed analogous informational lesioning experiments for N1 and N2. The results were a minimal change in performance when N1 was lesioned (98.59%) but a considerable decrease when N2 was lesioned (92.27%). Thus, the results indicate that N2 also plays a significant informational role with respect to $F$. In particular, it is likely that N2 accounts for some of the features in the plot of information gain for N3 (Figure 7.4), though further testing would be needed to determine exactly which of these features are due to N2 versus which are due to the sensors. However, in general the results make clear that, contrary to the view of one neuron capturing size information, information is in fact distributed both spatially—across different components of the system—and temporally—with different components carrying information at various times.

Let us now examine how the agent integrates information about first and second object size (abbreviated $S$). Rather than an exhaustive informational analysis, our primary interest here will be to compare with previous dynamical findings; thus, we will focus primarily on N1, which we know plays a critical role in triggering the agent's decision. To understand the information flow for each feature, we found it most useful to consider different fixed values of the other stimulus, instead of averaging over them. Thus, we first examined the flow of information for $F$ with different fixed values of $S$. An example is shown in Figure 7.7A for $S = 35$, with qualitatively similar results for other values of $S$. In examining these plots, we can ignore values occurring after time 110, which happen after the catch/avoid decision is made, and also the narrow band of high values occurring around size 35, which correspond to values of $F$ too similar to $S$ for the agent to discriminate. Thus, the predominant feature is a rapid increase in information about all sizes, followed by a sudden collapse just prior to the catch/avoid decision. For different values of $S$, this same feature is also present, but with the timing of the collapse occurring earlier or later depending on $S$. Moreover, examining the flow of information about $S$ for different fixed values of $F$ (Figure 7.7B), we find essentially the same pattern. N1 quickly gains information about $S$, then loses it all just before the response is made. Thus, somewhat strangely, although we know that N1 plays a crucial role in driving the agent's response, N1 suddenly loses all information about first and second object size just prior to the decision being made.

Of course, this result is not actually strange when we consider that what matters for

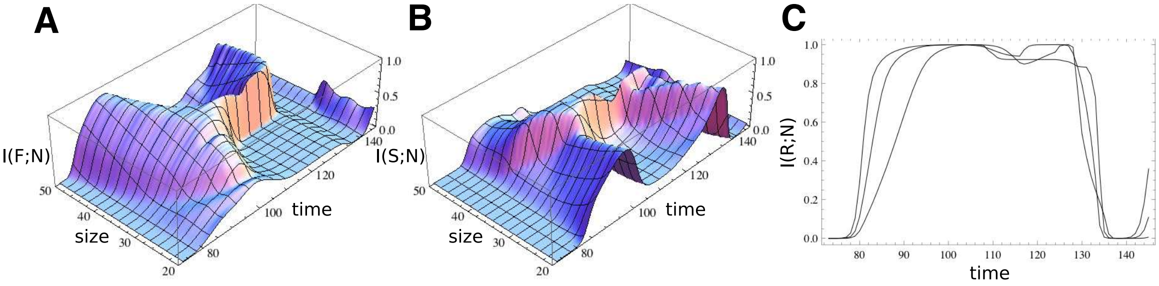


Figure 7.7: Flow of information about (a) first object size, (b) second object size, and (c) relative object size for N1 during the second half of the trial.

categorization is not the size of either object individually, but rather their relative size. To investigate relative size information, we can define a binary random variable $R$ corresponding to whether or not the second object is smaller than the first. The flow of information for $R$ is shown in Figure 7.7C for several different values of $S$. As expected, these plots show a rapid increase in relative size information for N1, reflecting the neural precursor to the agent's behavioral response. In addition, the slope of this increase can be seen to vary for different values of $S$, corresponding to the varying time of the agent's decision for different second object sizes. The timing of this increase in relative size information also coincides with the collapse in information about $F$ and $S$, so that collectively these features signal the integration of size information in N1. Interestingly, this timing also aligns with the timing of the underlying bifurcation that occurs to split the trajectories into catch and avoid bundles (recall the discussion in Section 6.5). Thus, here we observe an agreement between the dynamical and informational accounts of our relational agent, with each providing a different view of how its behavioral response is produced. Dynamically, this response is carried out through the timing of a bifurcation relative to a previous build-up of state, while informationally the response is captured by a sudden increase in size information for both objects, followed by a collapse in information about individual object size and a corresponding rise in information about relative size.

Using the techniques of information dynamics, we can also track how the relative size

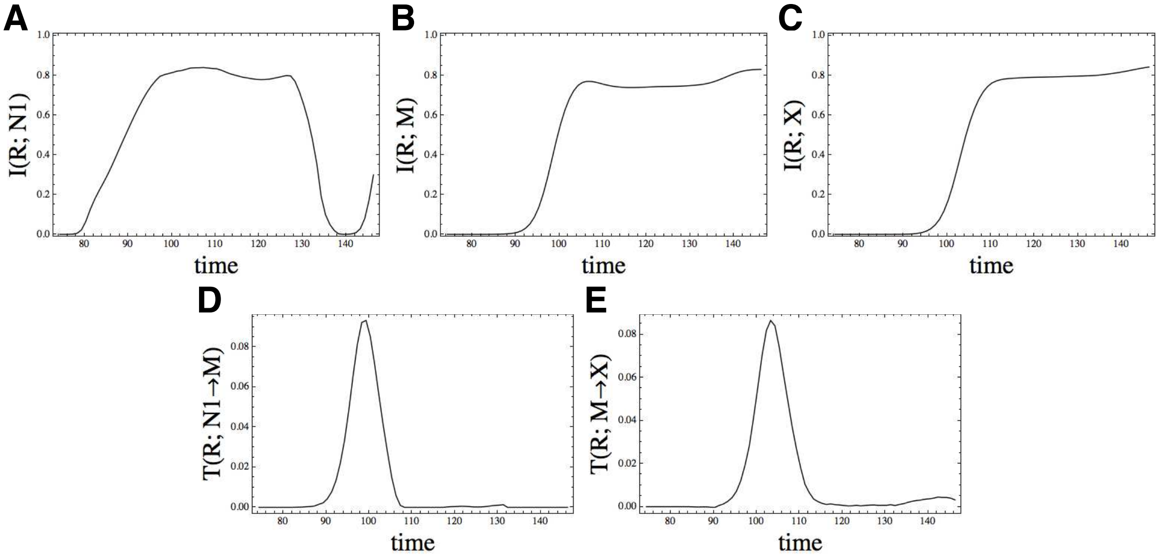


Figure 7.8: The flow of relative size information for (A) neuron N1, (B) the right motor neuron $M$, and (C) the agent's x-position $X$. The transfer of relative size information from $N1 \to M$ and from $M \to X$ are shown in (D) and (E), respectively.

information that builds up in N1 ultimately drives the agent's behavioral response. In particular, shortly after relative size information accumulates in N1 (Figure 7.8A), a similar pattern can be observed for the agent's right motor neuron (Figure 7.8B), followed in turn by a similar pattern for the agent's horizontal position (Figure 7.8C). This sequential build-up of relative size information in $N1$, in the right motor neuron ($M$), and finally in $X$, suggests a direct transfer of relative size information along the path $N1 \to M \to X$, where the high amount of information that ends up in $X$ is the information-theoretic manifestation of the agent's catch/avoid response. Indeed, calculating the information transfer from $N1 \to M$ and from $M \to X$ confirms that this is the case, with a high amount of transfer from $N1 \to M$ peaking around time $100$ (Figure 7.8D), followed by a similar peak in information transfer from $M \to X$ peaking around time $105$ (Figure 7.8E). Thus, from these plots we see that relative size information first builds up in $N1$, then flows to the right motor neuron and subsequently flows to $X$ where it is reflected in the agent's behavior.

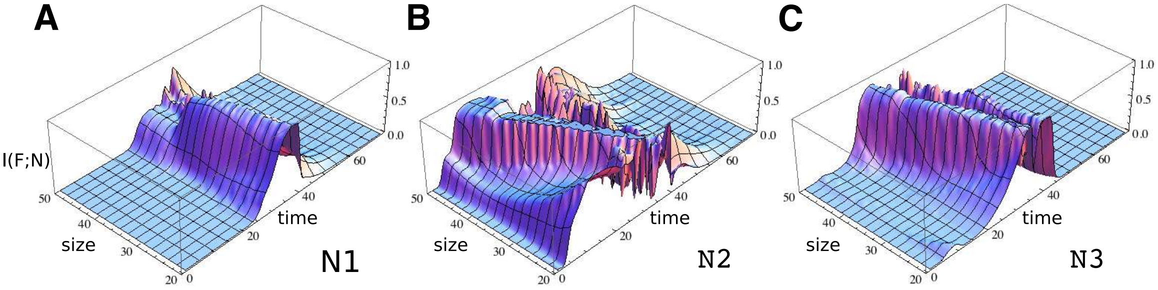
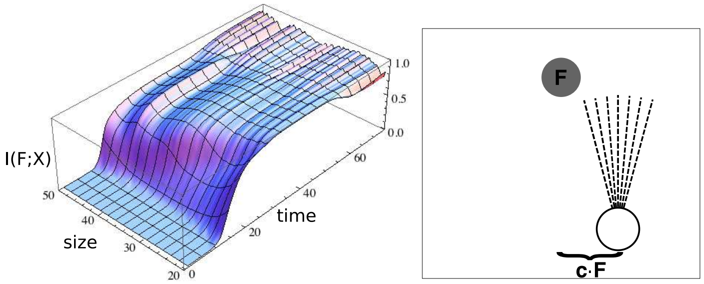


Figure 7.9: Flow of first object size information for neurons N1-N3 of the active agent.

## 7.3 Analysis of Active Relational Categorization

In this section, we explore the information dynamics of the best evolved active relational agent. Recall that, unlike the passive agent discussed in the previous section, this agent moves continuously during both the first and second object presentations (Figure 6.3(a)). Rather than perform a complete informational analysis of the active agent, here we focus on several features of its behavior that illustrate the various ways that embodiment can shape information dynamics.

The first feature has to do with how the active agent stores information about the size of the first object. As in our analysis of the passive agent, we can begin to address this question by examining the flow of first object size information for each of the agent's interneurons (Figure 7.9). And again, as for the passive agent, we find that all three interneurons carry significant amounts of size information at different times during the object's fall. However, unlike the passive agent, by the time the object concludes its fall and is removed from the environment, all of this information has been lost and the agent's interneurons are completely uninformative regarding first object size. In contrast, when we perform an identical information flow analysis for the agent's horizontal position, we find that its information content increases steadily as the object approaches and remains high as the object is removed (Figure 7.10(a)). Behaviorally, this result is produced by the agent positioning itself to the right of its starting location a distance proportional to the size of the



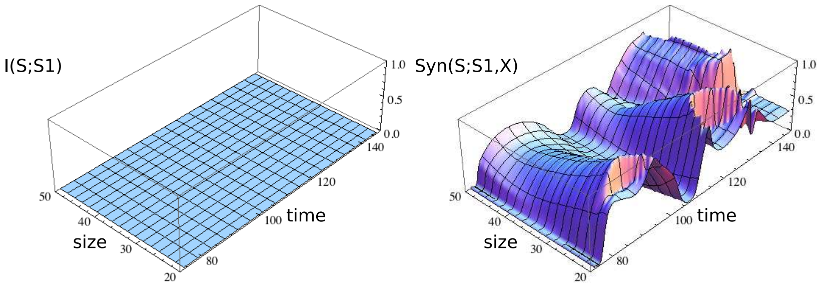
(a) (b)

Figure 7.10: Information offloading. (a) The flow of information about first object size for the agent's x-position. (b) Behaviorally, the increase in size information corresponds to the agent moving to the right a distance proportional to the size of the first object.

first object (Figure 7.10(b)). Thus, in this instance, the agent's brain demonstrably plays no part in storing information about object size[1], with this role instead performed by a component of the agent's body (or, more precisely, by a relationship between its body and its environment). These results present a crisp example of *information offloading*, which refers to when an embodied agent substitutes actions in the world for actions in the head, effectively offloading aspects of cognitive processes to the environment [106,119]. For example, laying out ingredients for a recipe in the order that they will be needed is an everyday instance of offloading. Although numerous studies have discussed the importance of this kind of offloading for embodied cognition, to our knowledge these results represent the first time it has been demonstrated within a rigorous information-theoretic framework.

This information offloading also impacts the information that the agent receives about the size of the second object. In particular, as a result of moving a variable distance dependent on the size of the first object, the second object will appear at differing locations in the agent's periphery. Informationally, this interaction between the agent's motion and where the second object appears is reflected in the flow of information about $S$ for the agent's interior sensors (those on the side where the second object falls). Figure 7.11(a) shows the flow of information about $S$ for sensor S1, the innermost sensor, revealing the surprising

[1] However, the agent's brain certainly plays a crucial role in getting information about object size into $X$.



(a)

(b)

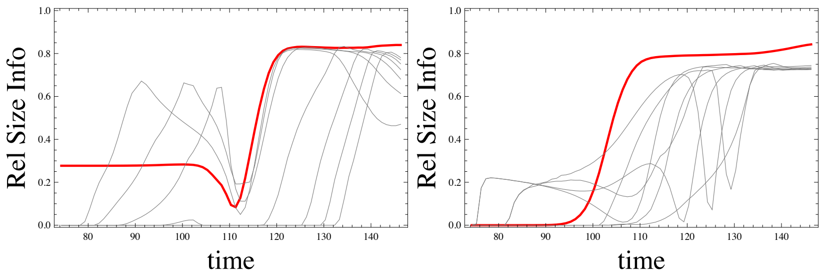
Figure 7.11: Self-structuring of second object size information. The flow of information about $S$ for (a) sensor S1 and (b) the synergy of S1 and the agent's horizontal position.

result that this sensor provides absolutely no information about $S$. Similar results are also observed for the other interior sensors, while the remaining sensors do not carry any information until well after the agent's catch or avoidance has been initiated. However, if instead one considers the flow of information provided by the *synergy* of S1 coupled with the agent's horizontal position (Figure 7.11(b)), we find that the sensor is indeed highly informative about the size of the second object, and likewise for the other interior sensors. This result can be viewed as a quantitative signature of *information self-structuring*, another hallmark of embodied information dynamics. Recall from the discussion in Section 3.4 that information self-structuring refers to an agent's use of its body to elicit and structure the information available from its sensory surfaces. In this case, it is purely through the synergistic interaction between the agent's sensors and the position of its body that the agent obtains any information at all about the size of the second object.

Finally, another example of information self-structuring, and one that is far more salient in the context of the categorization task, has to do with the flow of relative size information. Figure 7.12(a) shows the flow of information about $R$ for each of the active agent's sensors (gray lines) as well as for its horizontal position (red line)[2]. For comparison purposes, Figure 7.12(b) depicts the same information flows for the passive agent. Note that, for the active agent, several of its sensors provide substantial amounts of relative size information before such information becomes available in its horizontal position, i.e., before the agent initiates its catch/avoid response. In other words, prior to initiating its response, the agent has immediate access to relative size information from its sensors, thus enabling *direct perception* [73] of the higher-order variable $R$ and, as a result, vastly simplifying the categorization task. To envision how this might work, adopt the perspective of our active agent as the second object begins its fall. At this point in time, the agent is positioned to the right of the falling object, with the horizontal separation between the object and the agent proportional to the size of the first object. Consequently, if the agent has positioned itself in just the right way, the second object will either intersect its interior sensors if it is larger than the first, or not if it is smaller, and the agent's task reduces to simply attending to whether or not the activity of these sensors changes and responding accordingly[3]. In contrast, for the passive agent, none of its sensors provide substantial information about relative size until after the agent begins its response. This can be seen in Figure 7.12(b): none of the sensors begin increasing towards their peak information values until after information begins to rise for the agent's horizontal position. Indeed, it is impossible for the passive agent's sensors to gain any information about relative size, aside from the small amount accounted for in footnote 2, until the agent begins to move, allowing information

[2] Note that, curiously, the active agent's horizontal position carries some information about relative size even before the second object begins falling. At first blush, this result is strange indeed, as it suggests that the agent has some sort of *pre*cognitive ability! However, this seemingly odd result can be readily explained by the fact that object sizes are drawn uniformly from a fixed size interval of $[20, 50]$. Because of this, knowing the size of only one of the two objects often provides some information about their relative size: for instance, if the size of the first object is $50$ then the second object is guaranteed to be smaller, if the size is $45$ then the second object is highly likely to be smaller, and so forth. Thus, since the active agent's horizontal position starts out correlated with the size of the first object, it also provides some information about relative size. Similarly, the sensors of the passive agent provide some information about relative size when they begin to correlate with $S$, despite having no access to $F$ until the agent begins to move.

[3] Though the actual strategy employed by our active agent is slightly more complex, it appears to be similar in spirit to the one described here.



(a) active

(b) passive

Figure 7.12: Self-structuring of relative size information. The flow of relative size information for all sensors (gray lines) and the horizontal position (red line) of the (a) active and (b) passive agents.

about $F$ to flow along the path $N3 \rightarrow M \rightarrow X \rightarrow S$ and influence the sensors. Thus, in this instance, the benefits of information self-structuring are clearly demonstrated, with the active agent able to circumvent considerable cognitive demands simply by moving its body.

## 7.4 Discussion

One of the primary strengths of our information dynamics approach is that it applies naturally to situated and embodied aspects of behavior. As demonstrated in the previous section, techniques for analyzing information dynamics apply just as readily to bodily and environmental variables as they do to sensory and neural ones, and thus can be used to investigate interactions that span the brain-body and body-environment boundaries. In future work, we plan to apply these techniques to analyze other agents that exhibit interestingly embodied and extended solutions to cognitive tasks. For example, a previously evolved feature categorization agent [17] was found to repeatedly scan falling objects before deciding to catch or avoid them. We plan to investigate what benefits this scanning strategy confers over a more passive strategy, and thus perhaps to identify other advantages of information self-structuring. As a second example, we plan to explore how multiple agents in a shared environment use embodied interactions to transfer information to one another. In particular, we plan to explore this idea in a model of referential communication that we developed in previous work [253], in which one agent is required to communicate the locations of spatially distant targets to another agent solely through the use of embodied gestural interactions. In this case, the techniques of information dynamics can be used to track the flow of information not only through the sensors, neurons and body of a single agent, but also across the bodily divide separating two agents.

Another significant strength of our approach is that, by more fully extending information-theoretic techniques to the temporal domain, it opens up the possibility of exploring the relationship between informational and dynamical approaches to embodied systems. Although preliminary in nature, the analyses presented in Sections 6.5 and 7.2 point to some promising possibilities along these lines. For example, the two approaches were found to provide distinct yet compatible accounts for how the agent makes its categorical discrimination regarding relative object size. In informational terms, this was manifested by a sudden gain and then loss in information about both first and second object size in N1, followed by rapid gain in information about relative size, while dynamically this was explained by an appropriately timed bifurcation in the underlying dynamics. In addition, one can begin to see how the two approaches might complement one another. Using informational techniques, it was natural to explore how the agent extracts size information, a question that would be difficult to address or even formulate in dynamical terms. Conversely, the bifurcation identified via dynamical analysis helps to explain *why* the information flow in N1 exhibits a sudden collapse in information about particular sizes and increase in information about relative size.

More generally, it is our view that informational and dynamical approaches are likely to provide complementary insights, as a result of the unique perspectives and strengths that each approach affords. Dynamical tools are especially well-suited for characterizing the long-term behavior of systems, the stability of this behavior in response to perturbations, and the changes in long-term behavior that systems undergo when aspects of their structure are modified. In contrast, informational tools offer a view of how the behavior of a system is realized in terms of the specific interactions between its components, and may be especially useful for characterizing the non-autonomous and transient aspects of behavior that are tied to specific external features. As well as these differences, informational and dynamical ideas also share deep similarities, and it will undoubtedly be informative to explore this common ground as it relates to properties of brain-body-environment systems. For example, information gain and loss are closely related to the divergence or convergence of trajectories in phase space [134], which are characterized dynamically by the Lyapunov exponents [222]. In general, the divergence of trajectories leads to an increase in information, while convergence leads to its loss. Similarly, the limit sets and basin boundaries of a system also relate directly to its properties of convergence and divergence, and thus are likely to play a significant role in shaping the flow of information. However, the true test for all of these ideas will be to apply both dynamical and informational techniques to analyze concrete systems and to compare and contrast the resulting insights that each provides. The analysis presented here can be viewed as an initial step in this direction. Ultimately, such work may hopefully begin to reconcile the dynamical and information theoretic perspectives on intelligent agents which has generated so much controversy in recent years.

# 8

# Conclusion

The aim of this thesis has been to develop an information-theoretic framework for the analysis of embodied cognitive systems. This was done by first generalizing mutual information—one of the core concepts of information theory—to multivariate interactions with the new method of partial information decomposition. Second, we applied partial information decomposition to develop a general set of techniques for quantifying information dynamics, or how information about arbitrary features flows within, is gained and lost by, and is transferred between the components of a complex system. Finally, we applied these techniques to analyze specific examples of embodied cognitive systems. This analysis revealed how information about task relevant features was extracted and combined by the components of brain-body-environment systems, and illuminated some of the profound ways that embodiment can impact cognitive behavior.

In the next section, we summarize the main contributions of this thesis and discuss how it relates to, unifies and extends existing work. In the following section, we then conclude with several directions for future research that have opened up as a result of the work presented here.

## 8.1 Summary of Contributions

The core contributions of this thesis can be summarized in terms of the four key features of our framework identified in Chapter 3: its use of *multivariate*, *dynamic*, and *specific* informational measures to analyze *embodied* systems.

At the heart of our framework is the method of PI-decomposition, which provides a new way of quantifying the constituent elements of multivariate information. As discussed in Chapter 4, the existing work that most closely resembles PI-decomposition is the interaction information proposed by McGill [153], and the measures of synergy and redundancy used in neuroscience that derive from interaction information [25,131,169,196]. However, PI-decomposition significantly extends this previous work in that it exhaustively decomposes multivariate information into a collection of nonnegative terms, each of which can be interpreted unambiguously as a meaningful information quantity. In contrast, existing multivariate measures either confound different kinds of informational contributions, as in the case of interaction information, or group them together as a bulk quantity, as in the case of total correlation [247]. As we will discuss further in the next section, the fact that PI-decomposition can be used to decompose and examine the constituent parts of other informational measures also makes it a powerful tool for clarifying relationships between existing measures, as well as for developing new measures of interest.

In terms of dynamic measures, our framework also makes several significant contributions. First, the techniques described in Section 5.1 extend transfer entropy—a widely applied and influential measure of information transfer—in two significant ways, by decomposing it into state-dependent and state-independent components and by generalizing it to the multivariate case. Second, in Section 5.2, we showed that the same ideas could be broadened to encompass the flow of information about arbitrary variables, and in particular to include the flow of information about external stimulus features, as was our primary interest. In this way, the techniques developed in Section 5.2 provide a theoretical bridge between applications in neural coding, which are primarily concerned with quantifying information about particular stimuli, and dynamic informational measures, which have thus far been limited to quantifying intrinsic information flow. Finally, a significant feature of our application to embodied systems is that we unroll our dynamic measures over time to explore the detailed time course of information flow, in contrast with previous work which has focused solely on time-averaged measures. As discussed in Chapter 7, an important consequence of this kind of temporally-extended informational analysis is that it opens up the possibility to explore links between informational and dynamical perspectives on intelligent agents.

The concept of specific information featured prominently in the development of PI-decomposition described in Chaper 4. Indeed, the observation that different information sources may provide the same overall amount of information while providing information about different outcomes of a target variable $S$, and thus that information should be considered overlapping only when it is about the same outcome of $S$, was the key insight behind our redundancy measure $I_{\min}$. Specific information also played a central role in our analysis of embodied systems in Chapter 7, where it allowed us to quantify the flow of information about particular stimulus values. In this context, our use of specific information is closely related to applications in neural coding, where, as discussed in Section 3.3, it has been used to characterize the informativeness of neural responses as a function of stimulus condition [24,57,120,227]. However, our approach extends these previous applications in that we quantify specific information not only in neural but also in bodily and environmental variables, and we use specific information in conjunction with dynamic measures to track information as a function of both time and stimulus condition. As illustrated in Chapter 7, these extensions allow us to connect time- and stimulus-specific features of informational quantities (“waves” of information flow; “ridges” of information gain) with key aspects of agent-environment interactions (the timing of visual ray/object intersections).

Lastly, the work presented here also makes several noteworthy contributions with respect to the informational analysis of embodied systems. First, the relational categorization behavior explored here is considerably more complex than previously studied behaviors, which, as discussed in Section 3.4, have essentially been limited to simple forms of object tracking. Our analysis is also the first of its kind to use measures of specific information or extrinsic information flow, the first to explore the complete time course of information flow as opposed to time-averaged quantities, and the first to quantify information in bodily and environmental components as well as neural ones. As in previous studies, our analysis revealed information self-structuring as a striking feature of embodied cognitive systems, with the feedback from an agent's motion leaving characteristic informational signatures on its sensory stream. However, beyond simply demonstrating the occurrence of self-structuring, our analysis also highlighted its potential advantages for simplifying cognitively demanding tasks, something that has not been shown in previous studies. Finally, our analysis also revealed information offloading as another significant consequence of embodied action—the first time such a phenomenon has been identified in a rigorous information-theoretic manner—and demonstrated how offloading and self-structuring can be integrated seamlessly to produce intelligent behavior.

## 8.2 Future Work

Several directions for future work have been discussed throughout this thesis. Here we begin by summarizing these directions, and then conclude by identifying several other promising avenues of investigation.

In Chapter 4, we mentioned the need to devise efficient ways of calculating PI-terms, given that the number of such terms grows rapidly for increasing numbers of variables. Along these lines, we also noted that the lattice structure of these terms provides a useful constraint that could likely be exploited for their efficient calculation. In Chapter 5, we discussed how the relationship between intrinsic and extrinsic (or generalized) information flow might be used to unravel pathways of information exchange in neural networks or other complex systems. In particular, since intrinsic flow is likely an upper bound on extrinsic flow, measures of the former kind could first be used to identify significant transmitters and/or receivers of information, while the latter could then be used to decipher their incoming and outgoing transmissions. In addition, we noted that a measure of the transfer of information loss can be formulated by following a parallel line of reasoning to that in Section 5.2, and that such a measure may have useful interpretations in terms of attentional processes or something similar. Finally, in Chapter 7, we discussed the use of information dynamics to investigate other interestingly embodied and extended solutions to cognitive tasks, with the active scanning strategy used by a feature categorization agent [17] and embodied information transfer in a model of referential communication [253] being two tantalizing examples. Also, we mentioned the idea of exploring the relationship between informational and dynamical approaches to embodied systems, with the hope that one might harness the unique strengths and complementary aspects of each approach to obtain a more complete picture.

In addition to embodied and extended aspects of cognitive behavior, there are a number of other important issues raised by model agents that we hope to explore within our framework. For example, as with the relational agents, the feature categorization agent mentioned above must extract information from its stimuli, but differs from the relational agents in that, rather than storing the information, the agent must use it to produce an immediate behavioral response. Analyzing this agent will thus provide a complementary picture of how information can be extracted and used in an online fashion, as opposed to being stored for later use. Another interesting aspect of the feature categorization agent, and of other agents that perform a task involving selective attention [77,210,245], is that they must intelligently lose or suppress information. Categorical perception involves the suppression of differences between category members [93], and selective attention to certain features requires that other features be ignored. Thus, analyzing these agents could shed light on the mechanisms of information loss available to embodied agents, and could also clarify the cognitive significance of our measure of the transfer of information loss. Finally, a recently developed model of visually-guided interception [112] could be used to examine how an agent extracts information about object trajectories based on optic flow patterns and its own self-motion. A significant advantage of this model is that it can be directly connected to experimental studies of object interception in humans [52,60].

Other promising directions for future work involve the application of PI-decomposition and the techniques of information dynamics to other kinds of complex systems. Indeed, this is an area that is already being actively explored by ourselves and others. For example, in one recent study Timme and Beggs [228] used PI-decomposition, as well as several other multivariate information measures, to analyze Boolean logic gates and a back-propagation network trained to learn the same gates. The authors reported that, compared with other multivariate measures, PI-decomposition provided the most complete and accurate description of the interactions present in the logic gates, and also shed new light on the complexity of learning different rules via back-propagation. Specifically, the authors found that rules involving high amounts of synergy took longer to learn while rules involving high amounts of redundancy took less time to learn, a relationship that was obscured by other multivariate measures. In our own recent work, we have applied PI-decomposition to analyze the local information dynamics and emergent computational properties of cellular automata [62]. Building on earlier work by Lizier et al. [138–140], we developed a new set of spatiotemporal filters to more clearly separate out the emergent computational structures—background domains, particles, and particle collisions—that are associated with information storage, transfer, and modification, respectively, in cellular automata. With these filters, information storage is associated with highest-order redundancy, information transfer with unique information, and information modification with highest-order synergy. For future work, an obvious direction is to apply our techniques to analyze data from real neural systems. For example, PI-decomposition could be used to tease apart the synergistic and redundant contributions of neural encodings, with the aim of clarifying previous analyses, based on interaction information, which confound the two contributions [25, 131, 169, 196]. The techniques of information dynamics could also be used to perform more fine-grained analyses of functional brain networks, extending previous analyses based on transfer entropy [98]. Indeed, we are currently pursuing, in collaboration with experimental neuroscientists, investigations in both of these directions.

Finally, on the theoretical front, other promising research avenues involve using PI-decomposition to dissect and analyze the constituent terms of other informational measures. Investigations of this kind can both clarify relationships between existing measures and facilitate the development of new measures of interest. In this thesis, we saw two examples of this kind of dissection: first, when PI-decomposition was used to analyze interaction information, revealing the fact that it confounds synergistic and redundant terms (Figures 4.8 and 4.9); and second, when PI-decomposition was used to decompose transfer entropy into state-dependent and state-independent parts (Figure 5.1). As another example, in our work developing spatiotemporal filters for cellular automata described above, we began by decomposing the multivariate informational measures that were previously proposed as filters by Lizier et al. [62, 140]. By first decomposing these measures, we were then able to develop our new filters by either pulling out the constituent PI-terms that captured the desired computational properties, or by devising alternative PI-measures that better captured the properties of interest. In another recent study, James et al. [109] applied PI-decomposition to analyze the ways in which the past and future of a stochastic time series can contribute information about its present. Specifically, by decomposing the information that the past and future provide about the present into contributions that both the past and future provide redundantly, parts that are provided uniquely by the past or the future, and parts that are synergistically provided by both the past and future, the authors were able to connect the meanings of several dynamic informational measures and to draw out potential asymmetries between the past and future. In future work, similar PI-based analyses of other multivariate measures, such as those of neural complexity [230,232] and integrated information [12,231], could also be explored.

# A

# Brief Review of Lattice Theory

Here we review only the basic lattice-theoretic concepts needed for the results in Chapter 4. For a thorough treatment, see [44,85].

**Definition A.1.** *A pair* $\langle X, \leqslant \rangle$ *is a* partially ordered set *or* poset *if* $\leqslant$ *is a binary relation on* $X$ *that is reflexive, transitive and antisymmetric.*

**Definition A.2.** *Let* $Y \subseteq X$. *Then* $a \in Y$ *is a* maximal element *in* $Y$ *if for all* $b \in Y, a \leqslant b \Rightarrow a = b$. *A* minimal element *is defined dually. We denote the set of maximal elements of* $Y$ *by* $\overline{Y}$ *and the set of minimal elements by* $\underline{Y}$.

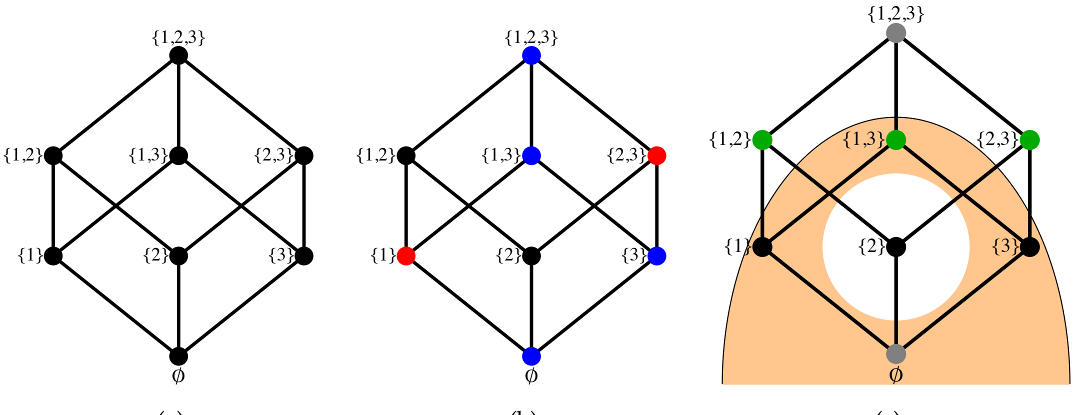
**Definition A.3.** *Let* $\langle X, \leqslant \rangle$ *be a poset, and let* $Y \subseteq X$. *An element* $x \in X$ *is an* upper bound *for* $Y$ *if for all* $y \in Y, y \leqslant x$. *A* lower bound *for* $Y$ *is defined dually.*

**Definition A.4.** *An element* $x \in X$ *is the* least upper bound *or* supremum *for* $Y$, *denoted* $\sup Y$, *if* $x$ *is an upper bound of* $Y$ *and for all* $y \in Y$ *and all* $z \in X, y \leqslant z$ *implies* $x \leqslant z$. *The* greatest upper bound *or* infimum *for* $Y$, *denoted* $\inf Y$, *is defined dually.*

**Definition A.5.** *A poset* $\langle X, \leqslant \rangle$ *is a* lattice *if, and only if, for all* $x, y \in X$ *both* $\inf\{x, y\}$ *and* $\sup\{x, y\}$ *exist in* $X$. *If* $\langle X, \leqslant \rangle$ *is a lattice, it is common to write* $x \wedge y$, *the* meet *of* $x$ *and* $y$, *and* $x \vee y$, *the* join *of* $x$ *and* $y$, *for* $\inf\{x, y\}$ *and* $\sup\{x, y\}$, *respectively. For* $Y \subseteq X$, *we use* $\bigwedge Y$ *and* $\bigvee Y$ *to denote the meet and join of all elements in* $Y$, *respectively.*

**Definition A.6.** *For* $a, b \in X$, *we say that* $a$ *is* covered by $b$ *(or* $b$ covers $a$*) if* $a < b$ *and* $a \leqslant c < b \Rightarrow a = c$. *The set of elements that are covered by* $b$ *is denoted by* $b^{-}$.

The classic example of a lattice is the power set of a set $X$ ordered by inclusion, denoted $\langle \mathcal{P}(X), \subseteq \rangle$. Lattices are naturally represented by Hasse diagrams, in which nodes



(a) (b) (c)

Figure A.1: Basic lattice-theoretic concepts. (A) Hasse diagram of the lattice $\langle \mathcal{P}(X), \subseteq \rangle$ for $X = \{1, 2, 3\}$. (B) An example of a chain (blue nodes) and an antichain (red nodes). (C) The top $\top$ and bottom $\perp$ are shown in gray. Green nodes correspond to $\{1, 2, 3\}^-$, the set of elements covered by $\{1, 2, 3\}$. The orange region represents $\downarrow \{1, 3\}$, the down-set of $\{1, 3\}$.

correspond to members of $X$ and an edge exists between elements $x$ and $y$ if $x$ covers $y$. Figure A.1(a) depicts the Hasse diagram for the lattice $\langle \mathcal{P}(X), \subseteq \rangle$ with $X = \{1, 2, 3\}$.

**Definition A.7.** *If $\langle X, \leqslant \rangle$ is a poset, $Y \subseteq X$ is a* chain *if for all $a, b \in Y$ either $a \leqslant b$ or $b \leqslant a$. $Y$ is an* antichain *if $a \leqslant b$ only if $a = b$.*

Figure A.1(b) shows examples of a chain and an antichain.

**Definition A.8.** *If there exists an element $\perp \in X$ with the property that $\perp \leqslant x$ for all $x \in X$, we call $\perp$ the* bottom element *of $X$. The* top element *of X, denoted by $\top$, is defined dually.*

**Definition A.9.** *For any $x \in X$, we define*

$$\downarrow x = \{y \in X : y \leqslant x\} \text{ and } \dot{\downarrow} x = \{y \in X : y < x\}$$

*where $\downarrow x$ and $\dot{\downarrow} x$ are called the* down-set *and* strict down-set *of $x$, respectively.*

Figure A.1(c) depicts top and bottom elements, covering relations, and down-sets.

# B

# Proof of Theorem 5.1

As defined in [235], a system has *perfect controllability* if, and only if, for any initial state $x$ and final state $x'$ there exists a controller state $c$ such that $p(x'|x,c) = 1$. We first prove that an equivalent definition of perfect controllability is that a system can be moved deterministically to any final state from any distribution of initial states.

**Lemma B.1.** *A system is perfectly controllable if, and only if, for any $x'$ there exists a distribution $p(c|x)$ such that $p(x') = 1$ for any distribution $p(x)$.*

*Proof.* If a system is perfectly controllable, we know that for a given $x'$ there exists at least one $c$ for each $x$ such that $p(x'|x,c) = 1$. Thus, we can choose

$$\operatorname{supp}(C|x) = \{c : p(x'|x,c) = 1\} \tag{B.1}$$

for each $x \in X$, which guarantees that $p(x') = 1$ for any distribution $p(x)$. As this is verified for any $x'$, this proves the direct part of the theorem.

Conversely, note that if $p(x') = 1$ for a given $x'$ and any distribution $p(x)$, it must be that

$$p(x'|x) = \sum_c p(x'|x,c)p(c|x) = 1 \tag{B.2}$$

for each $x \in X$, and thus for each $x$ there must be at least one $c$ for which $p(x'|x,c) = 1$. As this holds for any $x'$, the converse is proven. □

In order for a system to be perfectly controllable via open-loop control, we also require that the controller acts independently of the initial state, leading to the following definition.

**Definition B.1.** *A system has* perfect open-loop controllability *if, and only if, for any $x'$ there exists a distribution $p(c|x)$ such that $p(x') = 1$ for any distribution $p(x)$, and $I(X;C) = 0$.*

An alternative definition of perfect open-loop controllability is given by the following lemma.

**Lemma B.2.** *A system has perfect open-loop controllability if, and only if, for any $x'$ there exists a $c$ such that $p(x'|c) = 1$.*

*Proof.* If $I(X;C) = 0$, then $p(x')$ can be written as

$$p(x') = \sum_c p(c) \sum_x p(x)p(x'|x,c). \tag{B.3}$$

If we also have that $p(x') = 1$ for a given $x'$ and any distribution $p(x)$, then there must exist a $c$ for which $p(x'|x,c) = 1$ for all $x$, i.e., $p(x'|c) = 1$. This holds for any $x'$, so the direct part of the theorem is proven.

Conversely, if for any $x'$ there exists a $c$ such that $p(x'|c) = 1$, then for a given $x'$ we can choose

$$\mathrm{supp}(C) = \{c : p(x'|c) = 1\} \tag{B.4}$$

with $p(c) = p(c|x)$ for all $c$ and $x$, ensuring that $p(x') = 1$ and $I(X;C) = 0$. As this holds for any $x'$, the converse is proven. □

In [235], it is shown that an equivalent information-theoretic definition of perfect controllability is that there exists a distribution $p(c|x)$ such that each final state is reachable from each initial state, i.e.,

$$p(x'|x) \neq 0 \tag{B.5}$$

for all $x$ and $x'$ and that, for any distribution $p(x)$, $I(X';C|X)$ is maximal, i.e.,

$$H(X'|X,C) = 0 \tag{B.6}$$

so that

$$I(X';C|X) = H(X'|X) - H(X'|X,C) = H(X'|X). \tag{B.7}$$

Consequently, $I(X';C|X)$ is naturally interpreted as a system's degree of controllability, taking on its maximum value if, and only if, the system is perfectly controllable.

By the same reasoning, Theorem 5.1 establishes that the SITE from $C$ to $X'$ is a natural measure of a system's open-loop controllability, maximal exactly in the case of perfect open-loop control. In order to prove Theorem 5.1, we will need the following basic property of SDTE.

**Lemma B.3.** *The SDTE from $C$ to $X'$ is zero if, and only if, for each $x' \in X'$,*

$$p(x'|x,c) = p(x'|x)\ \forall x, c$$

*or*

$$p(x'|x,c) = p(x'|c)\ \forall x, c.$$

*Proof.*

$$\begin{aligned}
&\Pi(X';\{X,C\})\\
=&I(X';X,C) - I_{\max}(X';X,C)\\
=&\sum_{x'} p(x')\Big[I(X'=x';X,C) - \max\{I(X'=x';X), I(X'=x';C)\}\Big]\\
=&\sum_{x'} p(x')\min\{I(X'=x';C|X), I(X'=x';X|C)\}
\end{aligned}$$

where

$$I(X' = x'; C|X) = \sum_{x} p(x|x') D(p(c|x, x') \parallel p(c|x))$$

and $D(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence [36]. $D(\cdot \parallel \cdot)$ is nonnegative, so $\Pi(X'; \{X, C\}) = 0$ if, and only if, for each $x' \in X'$, $I(X' = x'; C|X) = 0$ or $I(X' = x'; X|C) = 0$. Furthermore, since $D(q \parallel r) = 0$ if, and only if, $q = r$, $\Pi(X'; \{X, C\}) = 0$ if, and only if, for each $x' \in X'$,

$$p(c|x, x') = p(c|x) \; \forall x, c$$

or

$$p(x|c, x') = p(x|c) \; \forall x, c$$

or, equivalently,

$$p(x'|x, c) = p(x'|x) \; \forall x, c$$

or

$$p(x'|x, c) = p(x'|c) \; \forall x, c.$$

□

Now we are in a position to prove Theorem 5.1. To do so, we will show that a system has perfect open-loop controllability if, and only if, there exists a distribution $p(c|x)$ such that Equations (B.5) and (B.6) are satisfied and there is no SDTE from $C$ to $X'$, i.e.,

$$\Pi(X'; \{X, C\}) = 0. \tag{B.8}$$

*Proof.* If a system is open-loop controllable, then for each $x'$ there exists a $c$ such that

$p(x'|c) = 1$. Choosing

$$\text{supp}(C) = \{c : p(x'|c) = 1\} \tag{B.9}$$

over all $x' \in X'$, with $p(c) = p(c|x)$ for all $c$ and $x$, ensures that $H(X'|X, C) \leq H(X'|C) = 0$ and $p(x'|x) \neq 0$. Also, the chosen distribution $p(c|x)$ ensures that $p(x'|x, c) = p(x'|c)\ \forall x, c$ so that $\Pi(X'; \{X, C\}) = 0$ by the preceding lemma. This proves the direct part of the theorem.

For the converse, note that $p(x'|x) \neq 0$ for a given $x'$ and $x$ means that there is at least one $c$ for which $p(x'|x, c) \neq 0$ and, since $H(X'|X, C) = 0$, we can further conclude that $p(x'|x, c) = 1$. If we also have that $\Pi(X'; \{X, C\}) = 0$, we know that, for each $x' \in X'$,

$$p(x'|x, c) = p(x'|x)\ \forall x, c$$

or

$$p(x'|x, c) = p(x'|c)\ \forall x, c$$

and thus that, for each $x' \in X'$,

$$\exists x, p(x'|x) = 1$$

or

$$\exists c, p(x'|c) = 1.$$

But it cannot be the case that $\exists x, p(x'|x) = 1$, since that would violate the reachability condition (Equation (B.5)), so we conclude that for each $x'$ there exists a $c$ such that $p(x'|c) = 1$. This proves the converse. □

# Bibliography

[1] M. Abeles, G. Hayon, and D. Lehmann. Modeling compositionality by dynamic binding of synfire chains. *Journal of Computational Neuroscience*, 17(2):179–201, 2004.

[2] S. Amari. Information geometry on hierarchical decomposition of stochastic interactions. *IEEE Transactions on Information Theory*, 47:1701–1711, 2001.

[3] P. O. Amblard and O. J. J. Michel. Measuring information flow in networks of stochastic processes. *Arxiv preprint arXiv:0911.2873*, 2009.

[4] P. O. Amblard and O. J. J. Michel. On directed information theory and granger causality graphs. *Journal of Computational Neuroscience*, 30(1):7–16, 2011.

[5] D. Anastassiou. Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, 3(83):1–8, 2007.

[6] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[7] B. B. Averbeck, P. E. Latham, and A. Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7:358–366, 2006.

[8] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich. Predictive information and explorative behavior of autonomous robots. *European Journal of Physics B*, 63:329–339, 2008.

[9] N. Ay, O. Eckhard, N. Bertschinger, and J. Jost. A unifying framework for complexity measures of finite systems. *Proceedings of the European Conference on Complex Systems*, page 80, 2006.

[10] N. Ay and D. Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(1):17–42, 2008.

[11] T. Bäck and H. P. Schwefel. An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1(1):1–23, 1993.

[12] D. Balduzzi and G. Tononi. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Computational Biology*, 4(6):e1000091, 2008.

[13] L. Barnett, C. L. Buckley, and S. Bullock. Neural complexity and structural connectivity. *Physical Review E*, 79(5):051914, 2009.

[14] L. Barnett, C. L. Buckley, and S. Bullock. A graph theoretic interpretation of neural complexity. *Arxiv preprint arXiv:1011.5334*, 2010.

[15] R. D. Beer. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3:469–509, 1995.

[16] R. D. Beer. Toward the evolution of dynamical neural networks for minimally cognitive behavior. *From Animals to Animats 4: Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, pages 421–429, 1996.

[17] R. D. Beer. The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243, 2003.

[18] R. D. Beer. Parameter space structure of continuous-time recurrent neural networks. *Neural Computation*, 18:3009–3051, 2006.

[19] R. D. Beer and J. C. Gallagher. Evolving dynamical neural networks for adaptive behavior. *Adaptive Behavior*, 1:91–122, 1992.

[20] R. D. Beer and P. L. Williams. Animals and animats: Why not both iguanas? *Adaptive Behavior*, 17:296–302, 2009.

[21] A. J. Bell. The co-information lattice. *Proceedings of ICA2003*, pages 921–926, 2003.

[22] W. Bialek, I. Nemenman, and N. Tishby. Complexity through nonextensivity. *Physica A: Statistical Mechanics and its Applications*, 302(1-4):89–99, 2001.

[23] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[24] A. Borst and F. E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2:947–958, 1999.

[25] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. de Ruyter van Steveninck. Synergy in a neural code. *Neural Computation*, 12(7):1531–1552, 2000.

[26] C. D. Brody, A. Hernandez, A. Zainos, and R. Romo. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebral Cortex*, 13(11):1196–1207, 2003.

[27] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7(5):456–461, 2004.

[28] D. A. Butts. How much information is associated with a particular stimulus? *Network*, 14:177–187, 2003.

[29] N. J. Cerf and C. Adami. Entropic Bell inequalities. *Physical Review A*, 55(5):3371–3374, 1997.

[30] P. Chanda, A. Zhang, D. Brazeau, L. Sucheston, J. L. Freudenheim, C. Ambrosone, and M. Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. *American Journal of Human Genetics*, 81(5):939–963, 2007.

[31] T. W. S. Chow and X. D. Li. Modeling of continuous time dynamical systems with input by recurrent neural networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 47(4):575–578, 2000.

[32] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[33] D. Cliff, I. Harvey, and P. Husbands. Explorations in evolutionary robotics. *Adaptive Behavior*, 2:73–110, 1993.

[34] L. Comtet. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Springer, 1974.

[35] R. G. Cook and E. A. Wasserman. Relational discrimination learning in pigeons. *Comparative cognition: Experimental explorations of animal intelligence*, pages 307–324, 2006.

[36] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[37] J. Crampton and G. Loizou. Two partial orders on the set of antichains. Research note, September 2000.

[38] J. Crampton and G. Loizou. The completion of a poset in a lattice of antichains. *International Mathematical Journal*, 1(3):223–238, 2001.

[39] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Physical Review E*, 55(2):1239–1242, 1997.

[40] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13:25, 2003.

[41] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

[42] I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

[43] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[44] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge Univ Press, 2nd edition, 2002.

[45] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

[46] H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[47] M. De Lucia, M. Bottaccio, M. Montuori, and L. Pietronero. Topological approach to neural complexity. *Physical Review E*, 71:016114, 2005.

[48] R. de Ruyter van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek. Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805, 1997.

[49] M. R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network*, 10:325–340, 1999.

[50] E. Di Paolo, J. Noble, and S. Bullock. Simulation models as opaque thought experiments. *Artificial Life VII: Proceedings of the Seventh International Conference on Artificial Life*, pages 497–506, 2000.

[51] E. A. Di Paolo, M. Rohda, and H. Iizuka. Sensitivity to social contingency of stability of interaction? modeling the dynamics of perceptual crossing. *New Ideas in Psychology*, 26:278–294, 2008.

[52] G. J. Diaz, F. Phillips, and B. R. Fajen. Intercepting moving targets: a little foresight helps a lot. *Experimental Brain Research*, 195(3):345–360, 2009.

[53] M. Ding, Y. Chen, and S. L. Bressler. Granger causality: Basic theory and application to neuroscience. *Handbook of Time Series Analysis*, pages 437–460, 2006.

[54] N. A. Dunn, S. R. Lockery, J. T. Pierce-Shimomura, and J. S. Conery. A neural network model of chemotaxis predicts functions of synaptic connections in the nematode caenorhabditis elegans. *Journal of Computational Neuroscience*, 17(2):137–147, 2004.

[55] W. Ebeling. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with lro. *Physica D: Nonlinear Phenomena*, 109(1):42–52, 1997.

[56] R. Eckhorn and B. Popel. Rigorous and extended application of information theory to the afferent visual system of the cat. i. basic concepts. *Kybernetik*, 16:191–200, 1974.

[57] R. Eckhorn and B. Pöpel. Rigorous and extended application of information theory to the afferent visual system of the cat. ii. experimental results. *Biological Cybernetics*, 17(1):7–17, 1975.

[58] J. Fagot, E. A. Wasserman, and M. E. Young. Discriminating the relation between relations: The role of entropy in abstract conceptualization by baboons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4):316, 2001.

[59] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

[60] B. R. Fajen and W. H. Warren. Behavioral dynamics of intercepting a moving target. *Experimental Brain Research*, 180:303–319, 2007.

[61] D. P. Feldman and J. P. Crutchfield. Measures of statistical complexity: Why? *Physics Letters A*, 238(4-5):244–252, 1998.

[62] B. Flecker, W. Alford, J. M. Beggs, P. L. Williams, and R. D. Beer. Partial information decomposition as a spatiotemporal filter. *Chaos*, in press.

[63] S. Frenzel and B. Pompe. Partial mutual information for coupling analysis of multivariate time series. *Physical Review Letters*, 99:204101, 2007.

[64] K.J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2:56–78, 1994.

[65] K. Funahashi and Y. Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801–806, 1993.

[66] W. R. Garner. *Uncertainty and Structure as Psychological Concepts*. Wiley, 1962.

[67] M. Gasser and E. Colunga. Where do relations come from? Technical Report 221, Indiana University, 1998.

[68] I. Gat and N. Tishby. Synergy and redundancy among brain cells of behaving monkeys. *Advances in Neural Information Processing Systems*, pages 111–117, 1999.

[69] T. J. Gawne and B. J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771, 1993.

[70] G. Gediga and I. Düntsch. On model evaluation, indices of importance, and interaction values in rough set analysis. In S. K. Pal, L. Polkowski, and A. Skowron, editors, *Rough-Neural Computing*. Physica Verlag, Heidelberg, 2003.

[71] D. Gentner. The development of relational category knowledge. *Building Object Categories in Developmental Time*, pages 245–275, 2005.

[72] D. Gentner and K. J. Kurtz. Relational categories. In *Categorization Inside and Outside the Laboratory*, pages 151–175. American Psychological Association, Washington, DC, 2005.

[73] J. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum, 1979.

[74] I. Gilboa and E. Lehrer. Global games. *International Journal of Game Theory*, 20(2):129–147, 1991.

[75] M. Giurfa, S. Zhang, A. Jenett, R. Menzel, and M. V. Srinivasan. The concepts of 'sameness' and 'difference' in an insect. *Nature*, 6831:930–933, 2001.

[76] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[77] E. Goldenberg, J. Garcowski, and R. D. Beer. May we have your attention: Analysis of a selective attention task. *From Animals to Animats 8: Proceedings of the Eighth International Confence on Simulation of Adaptive Behavior*, pages 49–56, 2004.

[78] M. B. Goldwater, A. B. Markman, and C. H. Stilwell. The empirical case for role-governed categories. *Cognition*, 118(3):359–376, 2010.

[79] I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics*, 34(3):911–934, 1963.

[80] B. Gourévitch and J. J. Eggermont. Evaluating information transfer between auditory cortical neurons. *Journal of Neurophysiology*, 97(3):2533–2543, 2007.

[81] M. Grabisch. Belief functions on lattices. *International Journal of Intelligent Systems*, 24(1):76–95, 2009.

[82] M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

[83] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[84] P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, 1986.

[85] G. A. Grätzer. *General Lattice Theory*. Birkhäuser, 2nd edition, 2003.

[86] S. Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1:17–61, 1988.

[87] Hu Guo-Dong. On the amount of information. *Teor. Veroyatnost. i Primenen*, 4:447–455, 1962. in Russian.

[88] G. S. Halford, W. H. Wilson, J. Guo, R. W. Gayler, J. Wiles, and J. E. M. Stewart. Connectionist implications for processing capacity limitations in analogies. *Advances in Connectionist and Neural Computation Theory*, 2:363–415, 1994.

[89] G. S. Halford, W. H. Wilson, and S. Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21(06):803–831, 1998.

[90] G. S. Halford, W. H. Wilson, and S. Phillips. Relational knowledge: the foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11):497–505, 2010.

[91] T. S. Han. Linear dependence structure of the entropy space. *Information and Control*, 29:337–368, 1975.

[92] T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46:26–45, 1980.

[93] S. Harnad. To cognize is to categorize: Cognition is categorization. In C. Lefebvre and H. Cohen, editors, *Handbook of Categorization in Cognitive Science*. Elsevier Press, 2005.

[94] I. Harvey, E. A. Di Paolo, R. Wood, M Quinn, and E. Tuci. Evolutionary robotics: a new scientific tool for studying cognition. *Artificial Life*, 11:79–98, 2005.

[95] R. Haschke and J. J. Steil. Input space bifurcation manifolds of recurrent neural networks. *Neurocomputing*, 64C:25–38, 2005.

[96] H. Hinrichs, H. J. Heinze, and M. A. Schoenfeld. Causal visual interactions as revealed by an information theoretic measure and fmri. *NeuroImage*, 31(3):1051–1060, 2006.

[97] K. Hlavácková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

[98] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Science USA*, 104(24):10240–10245, 2007.

[99] J. J. Hopfield. Neurons with graded response properties have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science USA*, 81:3088–3092, 1984.

[100] F. C. Hoppensteadt and E. M. Izhikevich. *Weakly connected neural networks*. Springer, Berlin, 1997.

[101] J. E. Hummel and K. J. Holyoak. Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review*, 104(3):427–466, 1997.

[102] J. E. Hummel and K. J. Holyoak. A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2):220, 2003.

[103] J. E. Hummel and K. J. Holyoak. Relational reasoning in a neurally plausible cognitive architecture. *Current Directions in Psychological Science*, 14(3):153, 2005.

[104] J. E. Hummel, K. J. Holyoak, C. Green, L. A. A. Doumas, D. Devnich, A. Kittur, and D. J. Kalar. A solution to the binding problem for compositional connectionism. *Compositional Connectionism in Cognitive Science: Papers from the AAAI Fall Symposium*, pages 31–34, 2004.

[105] Y. C. Hung and C. K. Hu. Chaotic communication via temporal transfer entropy. *Physical Review Letters*, 101(24):244102, 2008.

[106] E. Hutchins. *Cognition in the Wild*. MIT Press, Cambridge, MA, 1995.

[107] E. Izquierdo-Torres and I. Harvey. Learning on a continuum in evolved dynamical node networks. *Proceedings of Artificial Life X*, pages 507–512, 2006.

[108] A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions. *Arxiv preprint cs/0308002*, 2003.

[109] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *Arxiv preprint arXiv:1105.2988*, 2011.

[110] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of neural coding. *Journal of Computational Neuroscience*, 10(1):47–69, 2001.

[111] T. Jung, D. Polani, and P. Stone. Empowerment for continuous agent–environment systems. *Adaptive Behavior*, 19(1):16, 2011.

[112] D. Kadihasanoglu, R. D. Beer, and G. P. Bingham. The dependence of braking strategies on optical variables in an evolved model of visually-guided braking. *From Animals to Animats 11: Proceedings of the 11th International Conference on Simulation of Adaptive Behavior*, pages 555–564, 2010.

[113] A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1-2):43–62, 2002.

[114] K. Kaneko. Lyapunov analysis and information flow in coupled map lattices. *Physica D: Nonlinear Phenomena*, 23(1-3):436–447, 1986.

[115] A. Keinan. Controlled analysis of neurocontrollers with informational lesioning. *Philosophical Transactions of the Royal Society A*, 361(1811):2123–2144, 2003.

[116] C. Kemp and A. Jern. Abstraction and relational learning. *Advances in Neural Information Processing Systems*, 22:1–9, 2009.

[117] A. I. A. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, 1957.

[118] M. Kimura and R. Nakano. Learning dynamical systems by recurrent neural networks from orbits. *Neural Networks*, 11(9):1589–1599, 1998.

[119] D. Kirsh and P. Maglio. On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4):513–549, 1994.

[120] T. W. Kjaer, J. A. Hertz, and B. J. Richmond. Decoding cortical neuronal signals: network models, information estimation and spatial tuning. *Journal of Computational Neuroscience*, 1(1):109–139, 1994.

[121] G. J. Klir. *Uncertainty and Information*. Wiley Online Library, 2006.

[122] A. S. Klyubin, D. Polani, and C. L. Nehaniv. All else being equal be empowered. *Advances in Artificial Life*, pages 744–753, 2005.

[123] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Representations of space and time in the maximization of information flow in the perception-action loop. *Neural Computation*, 19(9):2387–2432, 2007.

[124] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PloS One*, 3(12):e4018, 2008.

[125] A. S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: A Universal Agent-Centric Measure of Control. *The 2005 IEEE Congress on Evolutionary Computation*, 1, 2005.

[126] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms, Volume 2*. Addison-Wesley, 1981.

[127] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[128] J. L. Krichmar, A. K. Seth, D. A. Nitz, J. G. Fleischer, and G. M. Edelman. Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions. *Neuroinformatics*, 3(3):197–222, 2005.

[129] K. J. Kurtz and O. Boukrina. Learning relational categories by comparison of paired examples. *Proceedings of the Conference of the Cognitive Science Society*, pages 756–761, 2004.

[130] O. Kwon and J. S. Yang. Information flow between composite stock index and individual stocks. *Physica A*, 387(12):2851–2856, 2008.

[131] P. E. Latham and S. Nirenberg. Synergy, redundancy, and independence in population codes, revisited. *Journal of Neuroscience*, 25(21):5195–5206, 2005.

[132] W. Li. On the relationship between complexity and entropy for markov chains and regular languages. *Complex Systems*, 5(4):381–399, 1991.

[133] H. Liang, M. Ding, and S. L. Bressler. Temporal dynamics of information flow in the cerebral cortex. *Neurocomputing*, 38–40:1429–1435, 2001.

[134] X. S. Liang and R. Kleeman. A rigorous formalism of information transfer between dynamical system components. II. Continuous flow. *Physica D*, 227(2):173–182, 2007.

[135] K. Lindgren and M. G. Nordahl. Complexity measures and cellular automata. *Complex Systems*, 2(4):409–440, 1988.

[136] J. T. Lizier, J. Heinzle, A. Horstmann, J. D. Haynes, and M. Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of Computational Neuroscience*, 30(1):85–107, 2011.

[137] J. T. Lizier, M. Prokopenko, I. Tanev, and A. Y. Zomaya. Emergence of glider-like structures in a modular robotic system. *Proceedings of Artificial Life X*, pages 366–377, 2008.

[138] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Detecting non-trivial computation in complex dynamics. *Proceedings of the 9th European Conference on Artificial Intelligence*, 4648:895–904, 2007.

[139] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 77:026110, 2008.

[140] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Information modification and particle collisions in distributed computation. *Chaos*, 20(3):037109, 2010.

[141] S. Lloyd. Measures of complexity: a nonexhaustive list. *IEEE Control Systems Magazine*, 21(4):7–8, 2001.

[142] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM Journal of Optimization*, 1:166–190, 1991.

[143] M. Lungarella and R. Pfeifer. Robots as cognitive tools: Information-theoretic analysis of sensory-motor data. *Proceedings of the 2nd International IEEE/RSJ Conference on Humanoid Robotics*, pages 245–252, 2001.

[144] M. Lungarella and O. Sporns. Information Self-Structuring: Key Principle for Learning and Development. *Proceedings of The 4th International Conference on Development and Learning*, pages 25–30, 2005.

[145] M. Lungarella and O. Sporns. Mapping Information Flow in Sensorimotor Networks. *PLoS Computational Biology*, 2(10), 2006.

[146] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[147] M. Madiman and P. Tetali. Information inequalities for joint distributions, with interpretations and applications. *IEEE Transactions on Information Theory*, 56(6):2699–2713, 2010.

[148] A. B. Markman and C. H. Stilwell. Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(4):329–358, 2001.

[149] R. Marschinski and H. Kantz. Analysing the information flow between financial time series. *European Physical Journal B*, 30(2):275–281, 2002.

[150] A. M. Mathai and P. N. Rathie. *Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications*. Wiley, 1975.

[151] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3):3096–3102, 2000.

[152] F. Matúš. Probabilistic conditional independence structures and matroid theory. *International Journal of General Systems*, 22(2):185–196, 1993.

[153] W. J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.

[154] R. Menzel and M. Giurfa. Cognitive architecture of a mini-brain: the honeybee. *Trends in Cognitive Sciences*, 5(2):62–71, 2001.

[155] P. Miller, C. D. Brody, R. Romo, and X. J. Wang. A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. *Cerebral Cortex*, 13(11):1208–1218, 2003.

[156] J. H. Moore, J. C. Gilbert, C. T. Tsai, F. T. Chiang, T. Holden, N. Barney, and B. C. White. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241(2):252–261, 2006.

[157] K. K. Nambiar, P. K. Varma, and V. Saroch. An axiomatic definition of Shannon's entropy. *Applied Mathematics Letters*, 5(4):45–46, 1992.

[158] N. S. Narayanan, E. Y. Kimchi, and M. Laubach. Redundancy and synergy of neuronal ensembles in motor cortex. *Journal of Neuroscience*, 25(17):4207–4216, 2005.

[159] C. Nehaniv, N. Mirza, and L. Olsson. Development via information self-structuring of sensorimotor experience and interaction. *50 years of Artificial Intelligence*, pages 87–98, 2007.

[160] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111, 2004.

[161] S. A. Neymotin, K. M. Jacobs, A. A. Fenton, and W. W. Lytton. Synaptic information transfer in computer models of neocortical columns. *Journal of Computational Neuroscience*, pages 1–16, 2011.

[162] S. Nolfi and D. Floreano. *Evolutionary Robotics*. MIT Press, 2000.

[163] K. Ono, M. Kudoh, and K. Shibuki. Relational discrimination learning between amplitude-modulated sounds in the rat. *Neuroscience Letters*, 342(3):171–174, 2003.

[164] L. Orlóci, M. Anand, and V. D. Pillar. Biodiversity analysis: issues, concepts, techniques. *Community Ecology*, 3(2):217–236, 2002.

[165] K. Otsuka, T. Ohtomo, A. Yoshioka, and J. Y. Ko. Collective chaos synchronization of pairs of modes in a chaotic three-mode laser. *Chaos*, 12:678–687, 2002.

[166] J. Pahle, A. K. Green, C. J. Dixon, and U. Kummer. Information transfer in signaling pathways: A study using coupled simulated and experimental data. *BMC Bioinformatics*, 9(1):139, 2008.

[167] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová. Synchronization as adjustment of information rates. *Physical Review E*, 63(4):46211, 2001.

[168] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[169] S. Panzeri, S. R. Schultz, A. Treves, and E. T. Rolls. Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society B*, 266(1423):1001–1012, 1999.

[170] F. Pasemann. Complex dynamics and the structure of small networks. *Network: Computation in Neural Systems*, 13:195–216, 2002.

[171] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[172] J. Pearl and G. Shafer. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[173] E. Pereda, R. Q. Quiroga, and J. Bhattacharya. Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2):1–37, 2005.

[174] R. Pfeifer and F. Iida. Embodied artificial intelligence: Trends and challenges. *Embodied Artificial Intelligence*, pages 629–629, 2004.

[175] R. Pfeifer, M. Lungarella, and F. Iida. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318(5853):1088, 2007.

[176] R. Pfeifer, M. Lungarella, O. Sporns, and Y. Kuniyoshi. On the information theoretic implications of embodiment–principles and methods. *50 years of Artificial intelligence*, pages 76–86, 2007.

[177] P. Phattanasri, H. J. Chiel, and R. D. Beer. The dynamics of associative learning in evolved model circuits. *Adaptive Behavior*, 15:377–396, 2007.

[178] T. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, 1995.

[179] D. Polani, O. Sporns, and M. Lungarella. How information and embodiment shape intelligent information processing. *50 years of Artificial Intelligence*, pages 99–111, 2007.

[180] M. Prokopenko, F. Boschetti, and A. J. Ryan. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28, 2009.

[181] M. Prokopenko, V. Gerasimov, and I. Tanev. Evolving spatiotemporal coordination in a modular robotic system. *Ninth International Conference on the Simulation of Adaptive Behavior*, pages 558–569, 2006.

[182] J. L. Puchalla, E. Schneidman, R. A. Harris, and M. J. Berry. Redundancy in the population code of the retina. *Neuron*, 46:493–504, 2005.

[183] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of Computational Neuroscience*, 30(1):17–44, 2011.

[184] M. Quinn. Evolving communication without dedicated communication channels. *Advances in Artificial Life: Proceedings of the Sixth European Conference on Artificial Life*, pages 357–366, 2001.

[185] R. Q. Quiroga and S. Panzeri. Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience*, 10:173–185, 2009.

[186] D. S. Reich, F. Mechler, and J. D. Victor. Independent and redundant information in nearby cortical neurons. *Science*, 294:2566–2568, 2001.

[187] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1999.

[188] P. Rodriguez, J. Wiles, and J. L. Elman. A recurrent neural network that learns to count. *Connection Science*, 11(1):5–40, 1999.

[189] R. Romo, C. D. Brody, A. Hernández, and L. Lemus. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473, 1999.

[190] S. M. Ross. *A First Course in Probability*. Prentice Hall, 8th edition, 2009.

[191] G. Rossi. Information functions and expectation. *Risk, Uncertainty and Decision*, pages 1–30, 2004.

[192] G. C. Rota. On the foundations of combinatorial theory I. Theory of Möbius functions. *Probability Theory and Related Fields*, 2(4):340–368, 1964.

[193] C. J. Rozell and D. H. Johnson. Examining methods for estimating mutual information in spiking neural systems. *Neurocomputing*, 65:429–434, 2005.

[194] E. L. Saldanha and M. E. Bitterman. Relational learning in the rat. *The American Journal of Psychology*, 64(1):37–53, 1951.

[195] C. Scheier and R. Pfeifer. Information theoretic implications of embodiment for neural network learning. *International Conference on Artificial Neural Networks*, pages 691–696, 1997.

[196] E. Schneidman, W. Bialek, and M. J. Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003.

[197] E. Schneidman, S. Still, M. J. Berry, and W. Bialek. Network information and connected correlations. *Physical Review Letters*, 91(23):238701, 2003.

[198] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

[199] D. W. Scott. Average shifted histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics*, 13:1024–1040, 1985.

[200] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley-Interscience, 1992.

[201] A. K. Seth. Causal connectivity of evolved neural networks during behavior. *Network: Computation in Neural Systems*, 16(1):35–54, 2005.

[202] A. K. Seth. Causal networks in simulated neural systems. *Cognitive Neurodynamics*, 2(1):49–64, 2008.

[203] A. K. Seth and G. M. Edelman. Environment and Behavior Influence the Complexity of Evolved Neural Networks. *Adaptive Behavior*, 12(1):5, 2004.

[204] A. K. Seth and G. M. Edelman. Distinguishing causal interactions in neural populations. *Neural Computation*, 19(4):910–933, 2007.

[205] C. E. Shannon. The bandwagon. *IRE Transactions in Information Theory*, 2(1):3, 1956.

[206] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[207] L. Shastri and V. Ajjanagadde. From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16:417–417, 1993.

[208] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[209] A. Singh and N. A. Lesica. Incremental mutual information: A new method for characterizing the strength and dynamics of connections in neuronal circuits. *PLoS Computational Biology*, 6(12):e1001035, 2010.

[210] A. Slocum, D. Downey, and R. D. Beer. Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. *From Animals to Animats 6*, pages 430–439, 2000.

[211] P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990.

[212] E. Sontag, A. Kiyatkin, and B. N. Kholodenko. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, 20(12):1877–1886, 2004.

[213] O. Sporns. *Networks of the Brain*. MIT Press, Cambridge, MA, 2011.

[214] O. Sporns, J. Karnowski, and M. Lungarella. Mapping Causal Relations in Sensorimotor Networks. *Proceedings the 5th International Workshop on Epigenetic Robotics*, 2006.

[215] O. Sporns and M. Lungarella. Evolving coordinated behavior by maximizing information structure. *Proceedings of Artificial Life X*, pages 3–7, 2006.

[216] O. Sporns and T. Pegors. Information-theoretical aspects of embodied artificial intelligence. *Embodied Artificial Intelligence*, pages 629–629, 2004.

[217] O. Sporns, G. Tononi, and G. M. Edelman. Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, 10:127–141, 2000.

[218] M. Staniek and K. Lehnertz. Symbolic transfer entropy. *Physical Review Letters*, 100(15):158101, 2008.

[219] R. P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press, 1997.

[220] B. Steudel and N. Ay. Information-theoretic inference of common ancestors. *Arxiv preprint arXiv:1010.5720*, 2010.

[221] B. Steudel, D. Janzing, and B. Schölkopf. Causal Markov condition for submodular information measures. *Arxiv preprint arXiv:1002.4020*, 2010.

[222] S. H. Strogatz. *Nonlinear Dynamics and Chaos*. Addison-Wesley, 1994.

[223] S. P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Physical Review Letters*, 80(1):197–200, 1998.

[224] M. Studený and J. Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. *Learning in Graphical Models*, pages 261–297, 1998.

[225] J. Szczepanski, M. Arnold, E. Wajnryb, J. M. Amigó, and M. V. Sanchez-Vives. Mutual information and redundancy in spontaneous communication between cortical neurons. *Biological Cybernetics*, pages 1–14, 2011.

[226] S. Takano. On 3-dimensional interaction information. *Proceedings of the Japan Academy*, 50(2):109–113, 1974.

[227] F. E. Theunissen and J. P. Miller. Representation of sensory information in the cricket cercal sensory system. ii. information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *Journal of Neurophysiology*, 66(5):1690–1703, 1991.

[228] N. Timme and J. M. Beggs. Towards resolving the multivariate information controversy. Unpublished manuscript.

[229] M. T. Tomlinson and B. C. Love. From pigeons to humans: Grouding relational learning in concrete examples. *Proceedings of the National Conference on Artificial Intelligence*, pages 199–204, 2006.

[230] G. Tononi, G. M. Edelman, and O. Sporns. Complexity and coherency: Integrating information in the brain. *Trends in Cognitive Science*, 2:474–484, 1998.

[231] G. Tononi and O. Sporns. Measuring information integration. *BMC Neurosci*, 4(1):31, 2003.

[232] G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Science USA*, 91(11):5033–5037, 1994.

[233] G. Tononi, O. Sporns, and G. M. Edelman. A complexity measure for selective matching of signals by the brain. *Proceedings of the National Academy of Science USA*, 93:3422–3427, 1996.

[234] H. Touchette and S. Lloyd. Information-theoretic limits of control. *Physical Review Letters*, 84(6):1156–1159, 2000.

[235] H. Touchette and S. Lloyd. Information-theoretic approach to the study of control systems. *Physica A*, 331(1-2):140–172, 2004.

[236] T. Tsujishita. On triple mutual information. *Advances in Applied Mathematics*, 16(3):269–274, 1995.

[237] E. Tuci, I. Harvey, and P. M. Todd. Using a net to catch a mate: Evolving ctrnns for the dowry problem. *From Animals to Animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, 2002.

[238] P. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, 2002.

[239] J. A. Vastano and H. L. Swinney. Information transport in spatiotemporal systems. *Physical Review Letters*, 60(18):1773–1776, 1988.

[240] V. Vedral. The role of relative entropy in quantum information theory. *Reviews of Modern Physics*, 74(1):197–234, 2002.

[241] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropya model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, pages 1–23, 2011.

[242] R. J. Vickerstaff and E. A. Di Paolo. Evolving neural models of path integration. *Journal of Experimental Biology*, 208:3349–3366, 2005.

[243] J. D. Victor. Approaches to information-theoretic analysis of neural activity. *Biological theory*, 1(3):302–316, 2006.

[244] R. Ward and R. Ward. Cognitive conflict without explicit conflict monitoring in a dynamical agent. *Neural Networks*, 19:1430–1436, 2006.

[245] R. Ward and R. Ward. Selective attention and control of action: Comparative psychology of an artificial, evolved agent and people. *Journal of Experimental Psychology: Human Perception and Performance*, 34:1165–1182, 2008.

[246] E. A. Wasserman, J. Fagot, and M. E. Young. Same–different conceptualization by baboons: The role of entropy. *Journal of Comparative Psychology*, 115(1):42, 2001.

[247] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

[248] D. Wiedemann. A computation of the eighth Dedekind number. *Order*, 8(1):5–6, 1991.

[249] P. L. Williams and R. D. Beer. Information dynamics of evolved agents. *From Animals to Animats 11: Proceedings of the Eleventh International Conference on the Simulation of Adaptive Behavior*, pages 38–49, 2010.

[250] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *Arxiv preprint cs/1004.2515*, 2010.

[251] P. L. Williams and R. D. Beer. Generalized measures of information transfer. *Arxiv preprint arXiv:1102.1507*, 2011.

[252] P. L. Williams, R. D. Beer, and M. Gasser. An embodied dynamical approach to relational categorization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 223–228, 2008.

[253] P. L. Williams, R. D. Beer, and M. Gasser. Evolving referential communication in embodied dynamical agents. *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 702–709, 2008.

[254] S. Wills. Relational learning in pigeons? *The Quarterly Journal of Experimental Psychology: Section B*, 52(1):31–52, 1999.

[255] L. Yaeger and O. Sporns. Evolution of neural structure and complexity in a computational ecology. *Artificial Life X*, 2006.

[256] L. Yaeger and O. Sporns. Passive and driven trends in the evolution of complexity. *Artificial Life XI*, 2008.

[257] L. Yaeger, O. Sporns, S. Williams, X. Shuai, and S. Dougherty. Evolutionary Selection

of Network Structure and Function. *Artificial Life XII: Proceedings of the Twelfth International Conference on the Simulation and Synthesis of Living Systems*, pages 313–320, 2010.

[258] R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Transactions on Information Theory*, 37(3):466–474, 1991.

[259] R. W. Yeung. *Information Theory and Network Coding*. Springer, 2008.

[260] M. E. Young and E. A. Wasserman. Entropy detection by pigeons: Response to mixed visual displays after same–different discrimination training. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(2):157, 1997.

[261] K. Zahedi, N. Ay, and R. Der. Higher coordination with less control—a result of information maximization in the sensorimotor loop. *Adaptive Behavior*, 18(3-4):338, 2010.

[262] Z. Zhang and R. W. Yeung. On characterization of entropy functions via information inequalities. *IEEE Transactions on Information Theory*, 44:1440–1452, 1998.

# Curriculum Vitae

My vitae.