Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Complexity as Causal Information Integration

by

Carlotta Langer and Nihat Ay

Preprint no.: 85 2020

Complexity as Causal Information Integration

Carlotta Langer[1] and Nihat Ay [1,2,3]

[1] Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[2] Leipzig University, Leipzig, Germany
[3] Santa Fe Institute, Santa Fe, USA

August 24, 2020

Abstract

Complexity measures in the context of the Integrated Information Theory of consciousness try to quantify the strength of the causal connections between different neurons. This is done by minimizing the KL-divergence between a full system and one without causal connections. Various measures have been proposed and compared in this setting. We will discuss a class of information geometric measures that aim at assessing the intrinsic causal influences in a system. One promising candidate of these measures, denoted by Φ_{CIS}, is based on conditional independence statements and does satisfy all of the properties that have been postulated as desirable. Unfortunately it does not have a graphical representation which makes it less intuitive and difficult to analyze. We propose an alternative approach using a latent variable which models a common exterior influence. This leads to a measure Φ_{CII}, Causal Information Integration, that satisfies all of the required conditions. Our measure can be calculated using an iterative information geometric algorithm, the em-algorithm. Therefore we are able to compare its behavior to existing integrated information measures.

keywords: Complexity; Integrated Information; Causality; Conditional Independence; em-Algorithm

1 Introduction

The theory of Integrated Information aims at quantifying the amount and quality of consciousness of a neural network. It was originally proposed by Tononi and went through various phases of evolution, starting with one of the first papers "Consciousness and Complexity" [27] in 1999 to "Consciousness as Integrated Information: a Provisional Manifesto" [26] in 2008 and IIT 3.0 [21] in 2014 to ongoing research. Although important parts of the methodology of this theory changed or got extended the two key concepts determining consciousness that virtually stayed fixed are "Information" and "Integration". Information refers to the number of different states a system can be in and Integration describes the amount to which the information is integrated among different parts of it. In order to determine to what extent a system integrates information, one divides it into smaller parts and calculates how much the split system differs from the full one. There are various ways to define a split system and the difference between them. Therefore, there exist different branches of complexity measures in the context of Integrated Information.

In detail we will measure the distance between the full and the split system using the KL-divergence as proposed in [5]. This framework was further discussed in [8]. Oizumi et al. [22] and Amari et al. [4] summarize these ideas and add a Markov condition and an upper bound to clarify what a complexity measure should satisfy. We will discuss these conditions in the next section. Additionally they introduce one measure that satisfies all of these requirements. This measure is described by conditional independence statements and will be denoted here by Φ_{CIS}. We will introduce Φ_{CIS} along with two other

existing measures, namely Stochastic Interaction Φ_{SI} [6] and Geometric Integrated Information Φ_G [1]. Although Φ_{CIS} fits perfectly in the proposed framework, this measure does not correspond to a graphical representation and it is therefore difficult to analyze the nature of the measured information flow.

The main purpose of this paper is to propose a more intuitive approach using a latent variable which models a common exterior influence. This leads to a new measure, which we call Causal Information Integration Φ_{CII}. This measure is specifically created to only measure the intrinsic causal influences and it satisfies all the required conditions postulated by Oizumi et al.

We discuss the relationship between the introduced measures in Section 2.0.2 and present a way of calculating Φ_{CII} by using an iterative information geometric algorithm, the em-algorithm described in Section 2.0.3. Utilizing this algorithm we are able to compare the behavior of Φ_{CII} to existing integrated information measures.

1.1 Integrated Information Measures

Measures corresponding to Integrated Information investigate the information flow in a system from a time t to $t+1$. This flow is represented by the connections from the nodes X_i in t to the nodes Y_i in $t+1$, $i \in \{1, \dots, n\}$ as displayed in Figure 1.

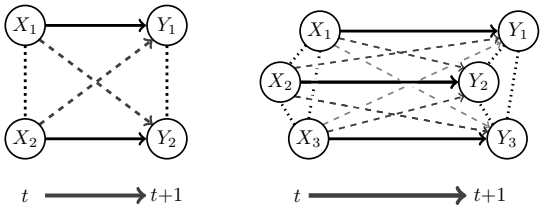


Figure 1: The fully connected system for $n = 2$ and $n = 3$.

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

$$X = (X_1, \dots, X_n) = (X_{1,t}, \dots, X_{n,t}), \qquad Y = (Y_1, \dots, Y_n) = (X_{1,t+1}, \dots, X_{n,t+1}), \qquad Z = (X, Y)$$

on a finite set $\mathcal{Z} \neq \emptyset$, which is the Cartesian product of the sample spaces of X_i $i \in \{1 \dots n\}$, denoted by $\mathcal{X}_i$

$$\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \bigtimes_{i=1}^{n} \mathcal{X}_i \times \bigtimes_{i=1}^{n} \mathcal{Y}_i.$$

Since the process is stationary and markovian we are able to restrict the discussion to one time step. Denote the complement of X_i in X by $X_{I\setminus\{i\}} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ with $I = \{1, \dots, n\}$. Corresponding to this notation $x_{I\setminus\{i\}} \in \mathcal{X}_{I\setminus\{i\}}$ describes the elementary events of $X_{I\setminus\{i\}}$. We will use the analogue notation in the case of Y and we will write $z \in \mathcal{Z}$ instead of $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The set of probability distributions on $\mathcal{Z}$ will be denoted by $\mathcal{P}(\mathcal{Z})$. Throughout this article we will restrict attention to strictly positive distributions.

The core idea of measuring Integrated Information is to determine how much the initial system differs from one in which no information integration takes place. The former will be called a "full" system, because we allow all possible connections between the nodes, and the latter will be called a "split" system. Graphical representations of the full systems for $n = 2, 3$ and their connections are depicted in Figure 1. In this article we are using graphs which describe the conditional independence structure of the corresponding sets of distributions. An introduction to those is given in Appendix A.

Following the concept introduced in [5], the difference between the measures corresponding to the full and split systems will be calculated by using the KL-divergence.

Definition 1 (Complexity)**.** Let $\mathcal{M}$ be a set of probability distributions on $\mathcal{Z}$ corresponding to a split system. Then we minimize the KL-divergence between $\mathcal{M}$ and the distribution of the fully connected system $\tilde{P}$ to calculate the complexity

$$\Phi_{\mathcal{M}} = \min_{Q \in \mathcal{M}} D_{\mathcal{Z}}(\tilde{P} \parallel Q) = \sum_{z \in \mathcal{Z}} \tilde{P}(z) \, log \frac{\tilde{P}(z)}{Q(z)}$$

Minimizing the KL-divergence with respect to the second argument is called m-projection or rI-projection. Hence we will call $P^\star$ with

$$P^\star = \arg\min_{Q \in \mathcal{M}} D_{\mathcal{Z}}(\tilde{P} \parallel Q)$$

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

The question remains how to define the split model $\mathcal{M}$. We want to measure the information that gets integrated between different nodes in different points in time. In Figure 1 these are the dashed connections, also called causal connections.

In order to ensure that these connections are removed in the split system, the authors of [22] and [4] argue that Y_j should be independent of X_i given $X_{I \setminus \{i\}}$, $i \neq j$ leading to the following property.

Property 1. A valid split system should satisfy the Markov condition

$$Q(X_i, Y_j \mid X_{I \setminus \{i\}}) = Q(X_i \mid X_{I \setminus \{i\}}) Q(Y_j \mid X_{I \setminus \{i\}}), \quad i \neq j \tag{1}$$

with $Q \in \mathcal{P}(\mathcal{Z})$. This can also be written in the following form

$$Y_j \perp\!\!\!\perp X_i | X_{I \setminus \{i\}}. \tag{2}$$

Now we take a closer look at the remaining connections. The dotted lines connect nodes belonging to the same point in time. These connections represent the common exterior or interior influences affecting the nodes leading to an undirected edge. Since we want to measure the amount of integrated information between t and $t+1$, the distribution in t, and therefore the connection between the X_is, should stay unchanged in the split system. The dotted connections between the Y_is play an important role in Property 2. For this property, we will consider the split system in which the solid and dashed connections are removed.

The solid arrows represent the influence of a node in t on itself in $t+1$ and removing these arrows, in addition to the causal connections, leads to a system with completely disconnected points in time as shown in the first row of Figure 3. The distributions corresponding to this split system are

$$\mathcal{M}_I = \{Q \in \mathcal{P}(\mathcal{Z}) | Q(z) = Q(x)Q(y), \forall z = (x, y) \in \mathcal{Z}\}$$

and the measure Φ_I is given by the mutual information $I(X;Y)$, which is defined in the following way

$$\Phi_I = I(X;Y) = \sum_{z \in \mathcal{Z}} P(x,y) \, log \left(\frac{P(x,y)}{P(x)P(y)} \right).$$

Since there is no information flow between the time steps Oizumi et al. argue in [22] that an integrated information measure should be bounded from above by the mutual information.

Property 2. The mutual information should be an upper bound for a measure for Integrated Information

$$\Phi_{\mathcal{M}} = \min_{Q \in \mathcal{M}} D_{\mathcal{Z}}(\tilde{P} \mid Q) \leqslant I(X;Y).$$

Oizumi et al. [22] and Amari et al. [4] state that this property is necessary and give the following two arguments. On the one hand this takes into account that the Y_is might have a common exterior influence that affects all the Y_is. This is symbolized by the additional node W in Figure 2 and this should not contribute to the value of Integrated Information between the different points in time.

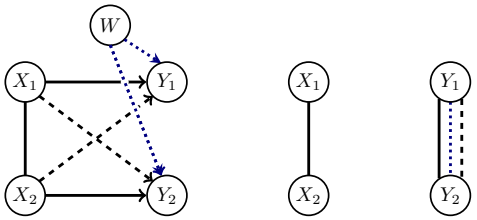


Figure 2: Interior and exterior influences on Y in the full and the split system corresponding to Φ_I.

On the other hand, we know that if the X_is are correlated, then the correlation is passed to the Y_is via the solid and dashed arrows. The question now is, how much of these correlations are causal and should therefore be measured. Kanwal et al. discuss this problem in [16]. They distinguish between intrinsic and extrinsic influences that cause the connections between the Y_is in the way displayed in Figure 2. By calculating the split system for Φ_I the edge between the Y_is might compensate for the solid arrows and common exterior influences, but also for the dashed, causal connections, as shown in Figure 2 on the right. Kanwal et al. analyze an example of a full system without a common exterior influence with the result that there are cases in which a measure that only removes the causal connections has a larger value than Φ_I. This is only possible if the undirected edge between the Y_is compensates also for a part of the causal connections. Hence Φ_I does not measure all the intrinsic causal influences. Therefore Kanwal et al. question the use of the mutual information as an upper bound.

Then again, we would like to contribute a different perspective. Admitting to Property 2 does not necessarily mean that the connections between the Y_is are fixed. It may merely mean that $\mathcal{M}_I$ is a subset of the set of split distributions. We will see that the measures Φ_{CIS} and Φ_{CII} do satisfy Property 2 in this way. Although the argument that Φ_I measures all the intrinsic influences is no longer valid, satisfying Property 2 is still desirable in general. Consider an initial system with the distribution $\tilde{P}(z) = \tilde{P}(x)\tilde{P}(y)$, $\forall z \in \mathcal{Z}$. This system has a common exterior influence on the Y_is and no connection between the different points in time. Since there is no information flow between the points in time, a measure for Integrated Information $\Phi_{\mathcal{M}}$ should be zero for all measures of this form. This is the case exactly when $\mathcal{M}_I \subseteq \mathcal{M}$, hence when Φ_I is an upper bound for $\Phi_{\mathcal{M}}$. In order to emphasize this point we propose a modified version of Property 2.

Property 3. The set $\mathcal{M}_I$ should be a subset of the split model $\mathcal{M}$ corresponding to the Integrated Information measure $\Phi_{\mathcal{M}}$. Then the inequality

$$\Phi_{\mathcal{M}} = \min_{Q \in \mathcal{M}} D_{\mathcal{Z}}(\tilde{P} \mid Q) \leqslant I(X;Y)$$

holds.

Note that the new formulation is stronger, hence Property 2 is a consequence of Property 3. Every measure discussed here that satisfies Property 2 fulfills also Property 3. Therefore we will keep referring to Property 2 in the following sections.

Figure 3 displays an overview over the different measures and whether they satisfy Properties 1 and 2. The first complexity measure that we are discussing does not fulfill Property 2. It is called Stochastic Interaction and was introduced by Ay in [5] in 2001, later published in [6]. Barrett and Seth discuss it in [9] in the context of Integrated Information. In [4] the corresponding model is called "fully split model".

The core idea is to allow only the connections among the random variables in t and additionally the connections between X_i and Y_i, meaning the same random variable in different points in time. The last ones correspond to the solid arrows in Figure 1. A graphical representation for $n = 2$ can be found in the first column of Figure 3.

Definition 2 (Stochastic Interaction)**.** The set of distributions belonging to the split model in the sense of Stochastic Interaction can be defined as

$$\mathcal{M}_{SI} = \left\{ Q \in \mathcal{P}(\mathcal{Z}) \mid Q(Y \mid X) = \bigotimes_{i=1}^{n} Q(Y_i \mid X_i) \right\}$$

and the complexity measure can be calculated as follows

$$\Phi_{SI} = \min_{Q \in M_{SI}} D_{\mathcal{Z}}(\tilde{P} \parallel Q) = \sum_{i=1}^{n} H(Y_i \mid X_i) - H(Y \mid X)$$

as shown in [6]. In the definition above, H denotes the conditional entropy

$$H(Y_i \mid X) = -\sum_{x \in \mathcal{X}} \sum_{y_i \in \mathcal{Y}_i} \tilde{P}(x, y_i) \, log \, \tilde{P}(y_i|x).$$

This does not satisfy Property 2 and therefore the corresponding graph is displayed only in the first column of Figure 3. Consider a setting without exterior influences, then this measure quantifies the strength of the causal connections alone and is therefore a reasonable choice for an Integrated Information measure. Accounting for an exterior influence that does not exist leads to a split system that compensates a part of the removal of the causal connections so that the resulting measure does not quantify all of the interior causal influences.

To force the model to satisfy Property 2, one can add the interaction between Y_i and Y_j which results in the measure Geometric Integrated Information [1].

Definition 3 (Geometric Integrated Information)**.** The graphical model corresponding to the graph in the second row and first column of Figure 3 is the set

$$\mathcal{M}_G = \left\{ P \in \mathcal{P}(\mathcal{Z}) | \exists f_1, \ldots, , f_{n+2} \in \mathbb{R}_+^{\mathcal{Z}} \text{ s.t. } P(z) = f_{n+1}(x) f_{n+2}(y) \prod_{i=1}^{n} f_i(x_i, y_i) \right\}$$

and the measure is defined as

$$\Phi_G = \min_{Q \in \mathcal{M}_G} D_{\mathcal{Z}}(\tilde{P} \parallel Q).$$

$\mathcal{M}_G$ is called the diagonally split model in [4]. This is not causally split in the sense that the corresponding distributions in general do not satisfy Property 1. It can be seen by analyzing the conditional independence structure of the graph as described in Appendix A. By introducing the edges between the Y_is as fixed, Φ_G might force these connections to be stronger than they originally are. A result of this might be that an effect of the causal connections gets atoned for by the new edge. We discussed this above in the context of Property 2 .

This measure has no closed form solution, but we are able to calculate the corresponding split system with the help of the iterative scaling algorithm, e.g. [12] Section 5.1.

The first measure that satifies both properties is called "Integrated Information" [22], its model is referred to by "Causally split model" in [4] and it is derived from the first property. Since we are able to define it using conditional independence statements, we will denote it by Φ_{CIS}. It requires Y_i to be independent of $X_{I \setminus \{i\}}$ given X_i.

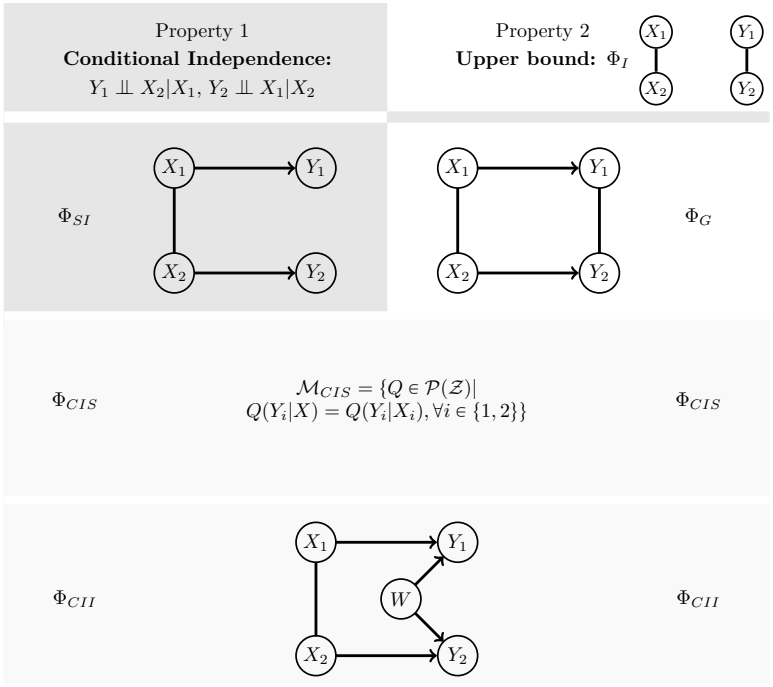


Figure 3: The different measures and their properties in the case of $n = 2$

Definition 4 (Integrated Information)**.** The set of distributions, that belongs to the split system corresponding to integrated information, is defined as

$$\mathcal{M}_{CIS} = \{Q \in \mathcal{P}(\mathcal{Z}) \mid Q(Y_i \mid X) = Q(Y_i \mid X_i), \text{ for all } i \in \{1, \dots, n\}\} \tag{3}$$

and this leads to the measure

$$\Phi_{CIS} = \min_{Q \in \mathcal{M}_{CIS}} D_{\mathcal{Z}}(\tilde{P} \parallel Q).$$

We write the requirements to the distributions in (3) as conditional independent statements

$$Y_i \perp\!\!\!\perp X_{I\setminus\{i\}} \mid X_i.$$

A detailed analysis of probabilistic independence statements can be found in [24]. Unfortunately, these conditional independence statements can not be encoded in terms of a chain graph in general. The definition of this measure arises naturally from Property 1 by applying the relation (1)

$$Q(X_i, Y_j \mid X_{I\setminus\{i\}}) = Q(X_i \mid X_{I\setminus\{i\}})Q(Y_j \mid X_{I\setminus\{i\}}), \quad i \neq j$$

to all pairs $i, j \in \{1, \dots, n\}$. This leads to

$$Q(Y_j|X) = Q(Y_j|X_j) \tag{4}$$

as shown in Appendix B.

Note that this implies that every model satisfying Property 1 is a submodel of $\mathcal{M}_{CIS}$. In order to show that Φ_{CIS} satisfies Property 1, we are going to rewrite the condition in Property 1 to

$$Q(Y_j|X) = Q(Y_j|X_{I\setminus\{i\}}).$$

The definition of $\mathcal{M}_{CIS}$ allows us to write

$$Q(Y_j|X) = Q(Y_j|X_j) = Q(Y_j|X_{I\setminus\{i\}})$$

for $Q \in \mathcal{M}_{CIS}$. Therefore Φ_{CIS} satisfies Property 1 and since $\mathcal{M}_I$ meets the conditional independence statements of Property 1 the relation $\mathcal{M}_I \subseteq \mathcal{M}_{CIS}$ holds and Φ_{CIS} fulfills Property 2.

In [22] Oizumi et al. derive an analytical solution for Gaussian variables, but there does not exist a closed form solution for discrete variables in general. Therefore they use Newton's method in the case of discrete variables.

Due to the lack of a graphical representation, it is difficult to interpret the causal nature of the elements of $\mathcal{M}_{CIS}$. In Example 1 we will see a type of model that is part of $\mathcal{M}_{CIS}$, but which challenges the notion of Integrated Information.

2 Causal Information Integration

Inspired by the discussion about extrinsic and intrinsic influences in the context of Property 2, we now utilize the notion of a common exterior influence to define the measure Φ_{CII}, which we call Causal Information Integration. Explicitly including a common exterior influence allows us to avoid the problems of a fixed edge between the Y_is discussed earlier. This leads to the graphs in Figure 4.

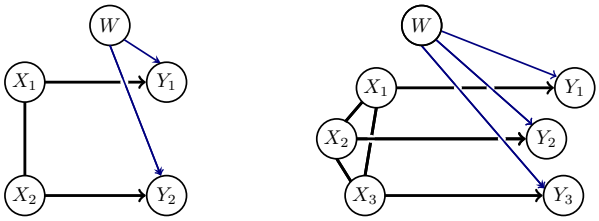


Figure 4: Split systems with exterior influences for $n = 2$ and $n = 3$.

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

$$P(z, w) = P(x) \prod_{i=1}^{n} P(y_i|x_i, w)P(w).$$

By marginalizing over the elements of $\mathcal{W}$ we get a distribution on $\mathcal{Z}$ defining our new model.

Definition 5 (Causal Information Integration)**.** The set of distributions belonging to the marginalized model for $|\mathcal{W}| = m$ is

$$\mathcal{M}^m_{CII} = \left\{ P \in \mathcal{P}(\mathcal{Z}) | \exists Q \in \mathcal{P}(\mathcal{Z} \times \mathcal{W}) : P(z) = \sum_{j=1}^{m} Q(x)Q(w_j) \prod_{i=1}^{n} Q(y_i|x_i, w_j) \right\}.$$

We will define the split model for Causal Integrated Information as

$$\mathcal{M}_{CII} = \bigcup_{m \in \mathbb{N}} \mathcal{M}^m_{CII}. \tag{5}$$

This leads to the measure

$$\Phi_{CII} = \min_{Q \in \mathcal{M}_{CII}} D_{\mathcal{Z}}(\tilde{P} \parallel Q).$$

In order to show that this measure satisfies the conditional independence statements in Property 1, we will calculate the conditional distributions $P(y_i|x_i)$ and $P(y_i|x)$ of

$$P(z) = \sum_w P(x) \prod_{j=1}^{n} P(y_j|x_j, w)P(w).$$

This results in

$$P(y_i|x_i) = \frac{\sum_{y_{I\setminus\{i\}}} \sum_{x_{I\setminus\{i\}}} \sum_w P(x) \prod_{i=j}^{n} P(y_j|x_j, w)P(w)}{P(x_i)} = \frac{\sum_{x_{I\setminus\{i\}}} \sum_w P(x)P(y_i|x_i, w)P(w)}{P(x_i)} = \sum_w P(y_i|x_i, w)P(w)$$

$$P(y_i|x) = \frac{\sum_{y_{I\setminus\{i\}}} \sum_w P(x) \prod_{i=j}^{n} P(y_j|x_j, w)P(w)}{P(x)} = \sum_w P(y_i|x_i, w)P(w)$$

for all $z \in \mathcal{Z}$. Hence $P(y_i|x_i) = P(y_i|x)$, Φ_{CII} satisfies Property 1 and the set of all such distributions is a subset of $\mathcal{M}_{CIS}$

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

We are able to represent the marginalized model by using the methods from [23]. Up to this point we have been using chain graphs. These are graphs consisting of directed and undirected edges such that there are no semi-directed cycles as described in Appendix A. In order to be able to gain a graph that represents the conditional independence structure of the marginalized model, we need the concept of chain mixed graphs (CMGs). In addition to the directed and undirected edges belonging to chain graphs, chain mixed graphs also have arcs $\leftrightarrow$. Two nodes connected by an arc are called spouses. The connection between spouses appears when we marginalize over a common influence, hence spouses do not have a directed information flow from one node to the other but are affected by the same mechanisms. The Algorithm 8 from [23] allows us to transform a chain graph with latent variables into a chain mixed graph that represents the conditional independence structures of the marginalized chain graph. Using this on the graphs in Figure 4 leads to the CMGs in Figure 5. Unfortunately, there exists no new factorization corresponding to the CMGs known to the authors.

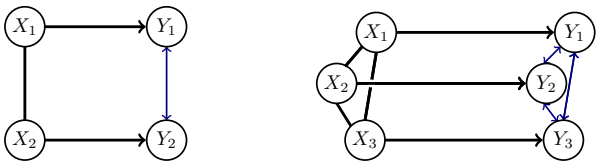


Figure 5: Marginalized Model for $n = 2$ and $n = 4$.

In order to show that Φ_{CII} satisfies Property 2, we will show that $\mathcal{M}_I$ is a subset of $\mathcal{M}_{CII}$. At first we will consider the following subset of $\mathcal{M}_{CII}$

$$\mathcal{M}_{CI}^m = \left\{ P \in \mathcal{P}(\mathcal{Z}) | \exists Q \in \mathcal{P}(\mathcal{Z} \times \mathcal{W}) : P(z) = \sum_{j=1}^{m} Q(x)Q(w_j) \prod_{i=1}^{n} Q(y_i|w_j) \right\}$$

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

where we remove the connections between the different stages, as shown in Figure 6.

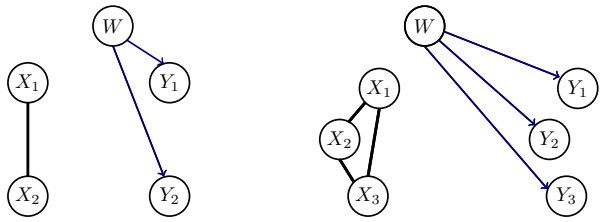


Figure 6: Submodels of the split models with exterior influences for $n = 2$ and $n = 3$.

Now X and Y are independent of each other

$$Q(z) = Q(x) \cdot Q(y)$$

with

$$Q(y) = \sum_w Q(w) \prod_{i=1}^{n} Q(y_i|w)$$

for $Q \in \mathcal{M}_{CI}$ and therefore we have $\mathcal{M}_{CI} \subseteq \mathcal{M}_I$. In order to gain equality it remains to show that $Q(Y)$ can approximate every distribution on $\mathcal{Y}$ if the state space of W is sufficiently large. These distributions are mixtures of discrete product distributions, where

$$\prod_{i=1}^{n} Q(y_i|w)$$

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

Theorem 2.1 (Theorem 1.3.1 from [20])**.** *Let q be a prime power. The smallest m for which any probability distribution on $\{1, \ldots, q\}$ can be approximated arbitrarily well as mixture of m product distributions is q^{n-1}.*

Universal approximation results like the theorem above may suggest that the models $\mathcal{M}_{CII}$ and $\mathcal{M}_{CIS}$ are equal. However we will present numerically calculated examples of elements belonging to $\mathcal{M}_{CIS}$, but not to $\mathcal{M}_{CII}$, even with an extremely large state space. We will discuss this matter further in Section 2.0.2.

In conclusion, Φ_{CII} satisfies Property 1 and 2.

2.0.1 Ground truth

The concept of an exterior influence suggests that there exists a ground truth in a larger model in which W is a visible variable. This is shown in Figure 7 on the right.

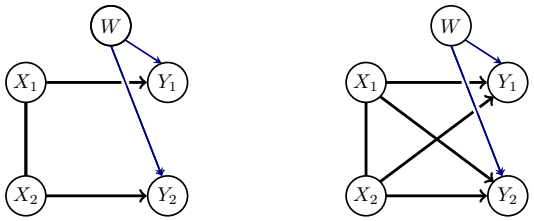


Figure 7: The graphs corresponding to $\mathcal{E}$ and $\mathcal{E}^f$ (right).

Assuming that we know the distribution of the whole model, we are able to apply the concepts discussed above to define an Integrated Information measure Φ_T on the larger space. This allows us to really only remove the causal connections as shown in Figure 7 on the left. Thus we can interpret Φ_T as the ultimate measure of Integrated Information, if the ground truth is available.

The set of distributions belonging to the larger, fully connected model will be called $\mathcal{E}^f$ and the set corresponding to the graph on the left of Figure 7 depicts the split system which will be denoted by $\mathcal{E}$.

$$\mathcal{E} = \bigcup_{m\in\mathbb{N}} \mathcal{E}^m,$$
$$\mathcal{E}^m = \left\{ P \in \mathcal{P}(\mathcal{Z}\times\mathcal{W}) \mid P(z,w) = P(x)\prod_{i=1}^{n} P(y_i|x_i,w)P(w),\ \forall (z,w)\in\mathcal{Z}\times\mathcal{W},\ |\mathcal{W}| = m \right\}$$
$$\mathcal{E}^f = \bigcup_{m\in\mathbb{N}} \mathcal{E}^{f,m},$$
$$\mathcal{E}^{f,m} = \left\{ P \in \mathcal{P}(\mathcal{Z}\times\mathcal{W}) \mid P(z,w) = P(x)\prod_{i=1}^{n} P(y_i|x,w)P(w),\ \forall (z,w)\in\mathcal{Z}\times\mathcal{W},\ |\mathcal{W}| = m \right\}$$

Note that $\mathcal{E}$ is the set of all the distributions that result in an element of $\mathcal{M}_{CII}$ after marginalization over $\mathcal{W}$

$$\mathcal{M}^m_{CII} = \left\{ P \in \mathcal{P}(\mathcal{Z}) | \exists Q \in \mathcal{E}^m : P(z) = \sum_{j=1}^{m} Q(x)Q(w_j)\prod_{i=1}^{n} Q(y_i|x_i,w_j) \right\}.$$

Calculating the KL-divergence between $P \in \mathcal{E}^f$ and $\mathcal{E}$ leads to the new measure.

Proposition 1. *Let $P \in \mathcal{E}^f$. Minimizing the KL-divergence between P and $\mathcal{E}$ leads to*

$$\begin{aligned}\Phi_T = \min_{Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) &= \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i P(y_i|x_i,w)} \\ &= \sum_i I(Y_i; X_{I\setminus\{i\}}|X_i, W).\end{aligned}$$

In the definition above $I(Y_i; X_{I\setminus\{i\}}|X_i, W)$ is the conditional mutual information defined by

$$I(Y_i; X_{I\setminus\{i\}}|X_i, W) = \sum_{y_i,x,w} P(y_i,x,w)\, log\, \frac{P(y_i, x_{I\setminus\{i\}}|x_i,w)}{P(y_i|x_i,w)P(x_{I\setminus\{i\}}|x_i,w)}.$$

It characterizes the reduction of uncertainty in Y_i due to $X_{I\setminus\{i\}}$ when W and X_i are given. Therefore this measure decomposes to a sum in which each addend characterizes the information flow towards one Y_i. Writing this as conditional independence statements, Φ_T is 0 if and only if

$$Y_i \perp\!\!\!\perp X_{I\setminus\{i\}} | \{X_i, W\}.$$

Ignoring W would lead exactly to Property 1. For a more detailed description of the conditional mutual information and its properties, see [11].

Additionally, by using that $W \perp\!\!\!\perp X$, we are able to split up the conditional mutual information into a part corresponding to the conditional independence statements of Property 1 and another conditional

mutual information.

$$\begin{aligned} I(Y_i; X_{I\setminus\{i\}}|X_i, W) &= \sum_{y_i,x,w} P(w)\, log\, \left(\frac{P(y_i, x_{I\setminus\{i\}}|x_i)}{P(y_i|x_i)P(x_{I\setminus\{i\}}|x_i)} \cdot \frac{P(y_i,x_i)P(x)P(y_i,x,w)P(x_i,w)}{P(y_i,x)P(x_i)P(y_i,x_i,w)P(x,w)} \right) \\ &= I(Y_i; X_{I\setminus\{i\}}|X_i) + \sum_{y_i,x,w} P(w)\, log\, \frac{P(y_i,x_i)P(x)P(y_i,x,w)P(x_i,w)}{P(y_i,x)P(x_i)P(y_i,x_i,w)P(x,w)} \\ &= I(Y_i; X_{I\setminus\{i\}}|X_i) + \sum_{y_i,x,w} P(w)\, log\, \frac{P(w, x_{I\setminus\{i\}}|y_i,x_i)}{P(w|y_i,x_i)P(x_{I\setminus\{i\}}|y_i,x_i)} \\ &= I(Y_i; X_{I\setminus\{i\}}|X_i) + I(W; X_{I\setminus\{i\}}|Y_i, X_i) \end{aligned}$$

Since the conditional mutual information is non-negative, Φ_T is 0 if and only if the conditional independence statements of Property 1 hold and additionally the reduction of uncertainty in W due to $X_{I\setminus\{i\}}$ given Y_i, X_i is 0.

In general, we do not know what the ground truth of our system is and therefore we have to assume that W is a hidden variable. This leads us back to Φ_{CII}. Since minimizing over all possible W might compensate a part of the causal information flow, Φ_{CII} is smaller or equal to the true value Φ_T.

Proposition 2. *The new measure Φ_T is an upper bound for Φ_{CII}*

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

Hence by assuming that there exists a common exterior influence, we are able to show that Φ_{CII} is bounded from above by the true value, that purely measures intrinsic influences.

2.0.2 Relationships between the different measures

Now we are going to analyze the relationship between the different measures $\Phi_{SI}, \Phi_G, \Phi_{CIS}$ and Φ_{CII}. We will start with Φ_G and Φ_{CII}. Previously we already showed that $\mathcal{M}_{CII}$ satisfies Property 1 and since Φ_G does not satisfy Property 1, we have

$$\mathcal{M}_G \nsubseteq \mathcal{M}_{CII}.$$

To evaluate the other inclusion, we will consider the more refined parametrizations of elements $P \in \mathcal{M}_{CII}$ and $Q \in \mathcal{M}_G$ as defined 6. These are

$$\begin{aligned} P(z) &= P(x) f_2(x_1, y_1) g_2(x_2, y_2) \sum_w P(w) f_1(w, y_1) f_3(x_1, y_1, w) g_1(w, y_2) g_3(x_2, y_2, w) \\ &= P(x) f_2(x_1, y_1) g_2(x_2, y_2) \phi(x_1, x_2, y_1, y_2) \\ Q(z) &= h_{n+1}(x) h_{n+2}(y) \prod_{i=1}^{n} h_i(y_i, x_i) \end{aligned}$$

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

$$\phi(x_1, x_2, y_1, y_2) = \sum_w P(w) f_1(w, y_1) f_3(x_1, y_1, w) g_1(w, y_2) g_3(x_2, y_2, w). \tag{6}$$

Since ϕ depends on more than Y_1 and Y_2, $P(z)$ does not factorize according to $\mathcal{M}_G$ in general. Hence $\mathcal{M}_{CII} \nsubseteq \mathcal{M}_G$ holds.

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

$$\begin{aligned} P(z) &= P(x) \tilde{f}_2(x_1, y_1) \tilde{g}_2(x_2, y_2) \sum_w P(w) \tilde{f}_1(w, y_1) \tilde{g}_1(w, y_2) \\ &= P(x) \tilde{f}_2(x_1, y_1) \tilde{g}_2(x_2, y_2) \tilde{\phi}(y_1, y_2) \end{aligned}$$

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

$$\tilde{\phi}(y_1, y_2) = \sum_w P(w)\tilde{f}_1(w, y_1)\tilde{g}_1(w, y_2).$$

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

In the next part we will discuss the relationship between $\mathcal{M}_{CII}$ and $\mathcal{M}_{CIS}$. The elements in $\mathcal{M}_{CII}$ satisfy the conditional independence statements of Property 1, therefore

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

Previously we have seen that making the state space of W large enough can approximate a distribution between the Y_is, see Theorem 2.1 . This seems to hint that doing so would lead to an equality between $\mathcal{M}_{CII}$ and $\mathcal{M}_{CIS}$, but based on numerically calculated examples, we have the following conjecture.

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

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

The following example strongly suggests this conjecture to be true.

Example 1. Consider the set of distributions that factor according to the graph in Figure 8

$$\mathcal{N}_{CIS}\{P \in \mathcal{P}(\mathcal{Z}) | P(z) = P(x_1)P(x_2)P(y_1|x_1, y_2)P(y_2)\}.$$

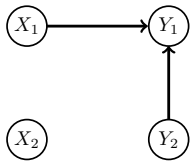


Figure 8: Graph of the model $\mathcal{N}_{CIS}$.

This model satisfies the conditional independence statements of Property 1 and is therefore a subset of the model $\mathcal{M}_{CIS}$. In this case X_1 and X_2 are independent of each other, hence from a causal perspective the influence of Y_2 on Y_1 should be purely external. Therefore we try to model this with a subset of $\mathcal{M}_{CII}$

$$\begin{aligned} \mathcal{N}_{CII} &= \bigcup_{m \in \mathbb{N}} \mathcal{N}_{CII}^m, \\ \mathcal{N}_{CII}^m &= \{P \in \mathcal{P}(\mathcal{Z}) | \exists Q \in \mathcal{P}(\mathcal{Z} \times \mathcal{W}) : P(z) = Q(x_1)Q(x_2) \sum_{j=1}^{m} Q(y_1|x_1, w_j)Q(y_2|w_j)Q(w_j)\} \end{aligned} \tag{7}$$

and this corresponds to Figure 9.

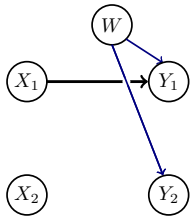


Figure 9: Graph of the model $\mathcal{N}_{CII}$.

Using the em-algorithm described in Section 2.0.3 we took 500 random elements of $\mathcal{N}_{CIS}$ and calculated the closest element of $\mathcal{N}_{CII}$ by using the minimum Kl-divergence of 50 different random input distributions in each run. The results are displayed in Table 1.

$\lvert\mathcal{W}\rvert$	minimum	maximum	arithmetic mean
2	0.011969035529826939	0.5028091152589176	0.15263592877594967
3	0.021348311360946	0.5499395859771526	0.1538653506807848
4	0.014762084688030863	0.3984635189946462	0.15139198568055212
8	0.017334311629729246	0.4383731978333986	0.15481967618112732
16	0.024306996171092318	0.4238222051787452	0.1490336847067273
300	0.016524177216064712	0.47733473380366764	0.15493896625208842

Table 1: The results of the em-algorithm between $\mathcal{N}_{CIS}$ and $\mathcal{N}_{CII}$

If we trust the generated results, this would imply that the influence from Y_2 to Y_1 is not purely external, but that there suddenly develops an internal influence in timestep $t+1$ that did not exist in timestep t. This situation should not occur in the context of Integrated Information.

Further examples which hint towards $\mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS}$ can be found in Section 2.1.2.

Adding the hidden variable W seems not to be sufficient to approximate elements of $\mathcal{M}_{CIS}$. Now the question naturally arises whether there are other exterior influences that need to be included in order to be able to approximate $\mathcal{M}_{CIS}$. We will explore this thought by starting with the graph corresponding to the split model $\mathcal{M}_{SI}$, depicted in Figure 10 on the left. In the next step we add hidden vertices and edges to the graph in a way such that the whole graph is still a chain graph. An example for a valid hidden structure is given in Figure 10 in the middle. Since we are going to marginalize over the hidden structure, it is only important how the visible nodes are connected via the hidden nodes. In the case of the example in Figure 10 we have a directed path from X_1 to X_2 going through the hidden nodes. Therefore we are able to reduce the structure to a gray box shown on the right in Figure 10.

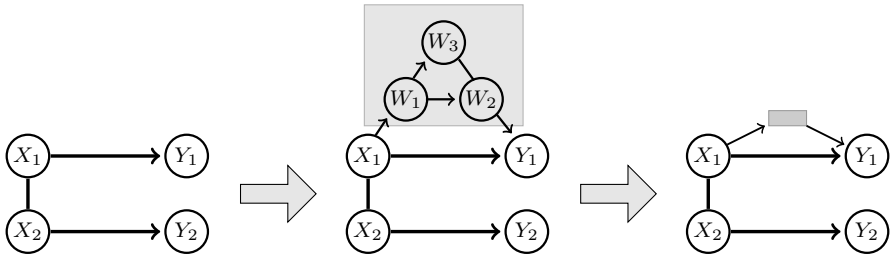


Figure 10: Example of an exterior influence on the initial graph.

Using the Algorithm 8 mentioned earlier that converts a chain graph with hidden variables to a chain mixed graph reflecting the conditional independence structure of the marginalized model, we gain that by marginalizing we would create a directed edge from X_1 to X_2. Seeing that this directed edge already existed, the resulting model now is a subset of $\mathcal{M}_{SI}$ and therefore does not approximate $\mathcal{M}_{CIS}$.

Following this procedure we are able to show that adding further hidden nodes and subgraphs of hidden nodes does not lead to a chain mixed graph belonging to a model that satisfies the conditional independence statements of Property 1 and strictly contains $\mathcal{M}_{CII}$.

Theorem 2.2. *It is not possible to create a chain mixed graph corresponding to a model $\mathcal{M}$, such that its distributions satisfy Property 1 and $\mathcal{M}_{CII} \subsetneq \mathcal{M}$, by introducing a more complicated hidden structure to the graph of $\mathcal{M}_{SI}$.*

In conclusion, assuming that Conjecture 1 holds, we have the following relations among the different presented models.

$$\mathcal{M}_I \subsetneq \mathcal{M}_G$$
$$\mathcal{M}_I \subsetneq \mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS}$$
$$\mathcal{M}_{SI} \subsetneq \mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS}$$

A sketch of the inclusion properties among the models is displayed in Figure 11.

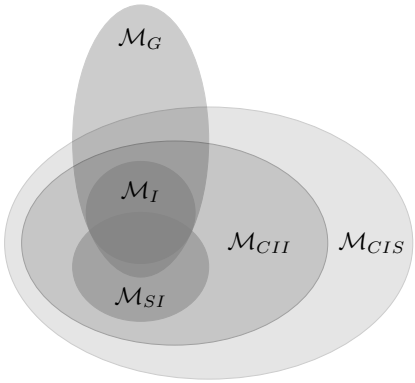


Figure 11: Sketch of the relationship between the manifolds corresponding to the different measures.

Every set that lies inside $\mathcal{M}_{CIS}$ satisfies Property 1 and every set that completely contains $\mathcal{M}_I$ fulfills Property 2.

2.0.3 em-Algorithm

The calculation of the measure Φ_{CII} can be done by the em-algorithm, a well known information geometric algorithm. It was proposed by Csiszár and Tusnády in 1984 in [13] and its usage in the context of neural networks with hidden variables was described for example by Amari et al. in [3]. The expectation-maximization EM-algorithm [14] used in statistics is equivalent to the em-algorithm in many cases, including this one, as we will see below. A detailed discussion of the relationship of these algorithms can be found in [2].

In order to calculate the distance between the distribution $\tilde{P}$ and the set $\mathcal{M}_{CII}$ on $\mathcal{Z}$ we will make use of the bigger space of distributions on $\mathcal{Z} \times \mathcal{W}$, $\mathcal{P}(\mathcal{Z} \times \mathcal{W})$. Let $\mathcal{M}_{W|Z}$ be the set of all distributions on $\mathcal{Z} \times \mathcal{W}$ that have $\mathcal{Z}$-marginals equal to the distribution of the whole system $\tilde{P}$

$$\begin{aligned}\mathcal{M}_{W|Z} &= \left\{ P \in \mathcal{P}(\mathcal{Z} \times \mathcal{W}) \mid P(z) = \tilde{P}(z),\, \forall z \in \mathcal{Z} \right\} \\ &= \left\{ P \in \mathcal{P}(\mathcal{Z} \times \mathcal{W}) \mid P(z,w) = \tilde{P}(z)P(w|z),\, \forall (z,w) \in \mathcal{Z} \times \mathcal{W} \right\}.\end{aligned}$$

This is an m-flat submanifold since it is linear w.r.t $P(w|z)$.

The second set that we are going to use is the set $\mathcal{E}$ of distributions that factor according to the split model including the common exterior influence. We have seen this set before in Section 2.0.1.

$$\begin{aligned}\mathcal{E} &= \bigcup_{m \in \mathbb{N}} \mathcal{E}^m \\ \mathcal{E}^m &= \left\{ P \in \mathcal{P}(\mathcal{Z} \times \mathcal{W}) \mid P(z,w) = P(x) \prod_{i=1}^{n} P(y_i|x_i,w)P(w),\, \forall (z,w) \in \mathcal{Z} \times \mathcal{W},\, |\mathcal{W}| = m \right\}\end{aligned} \tag{8}$$

This set is in general not e-flat, but we will show that there is a unique m-projection to it. We are able to use these sets instead of $\tilde{P}$ and $\mathcal{M}_{CII}$ because of the following result.

Theorem 2.3 (Theorem 7 from [3])**.** *The minimum divergence between $\mathcal{M}_{W|Z}$ and $\mathcal{E}$ is equal to the minimum divergence between $\tilde{P}$ and $\mathcal{M}_{CII}$ in the visible manifold*

$$\min_{P\in\mathcal{M}_{W|Z},Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) = \min_{\tilde{Q}\in\mathcal{M}_{CII}} D_{\mathcal{Z}}(\tilde{P} \parallel \tilde{Q}).$$

Proof of Theorem 2.3. Let $P, Q \in \mathcal{P}(\mathcal{Z}\times\mathcal{W})$, using the chain-rule for KL-divergence leads to

$$D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) = D_{\mathcal{Z}}(P \parallel Q) + D_{\mathcal{W}|\mathcal{Z}}(P \parallel Q)$$

with

$$D_{\mathcal{W}|\mathcal{Z}}(P \parallel Q) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w) log \frac{P(w|z)}{Q(w|z)}.$$

This results in

$$\begin{aligned}
\min_{P\in\mathcal{M}_{W|Z},Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) &= \min_{P\in\mathcal{M}_{W|Z},Q\in\mathcal{E}} \left\{D_{\mathcal{Z}}(P \parallel Q) + D_{\mathcal{W}|\mathcal{Z}}(P \parallel Q)\right\} \\
&= \min_{P\in\mathcal{M}_{W|Z},Q\in\mathcal{E}} \left\{D_{\mathcal{Z}}(\tilde{P} \parallel Q) + D_{\mathcal{W}|\mathcal{Z}}(P \parallel Q)\right\} \\
&= \min_{Q\in\mathcal{M}_{CII}} D_{\mathcal{Z}}(\tilde{P} \parallel Q).
\end{aligned}$$

□

The em-algorithm is an iterative algorithm that first performs an e-projection to $\mathcal{M}_{W|Z}$ and then an m-projection to $\mathcal{E}$ repeatedly. Let $Q_0 \in \mathcal{E}$ be an arbitrary starting point and define P_1 as the e-projection of Q_0 to $\mathcal{M}_{W|Z}$

$$P_1 = \arg\min_{P\in\mathcal{M}_{W|Z}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q_0).$$

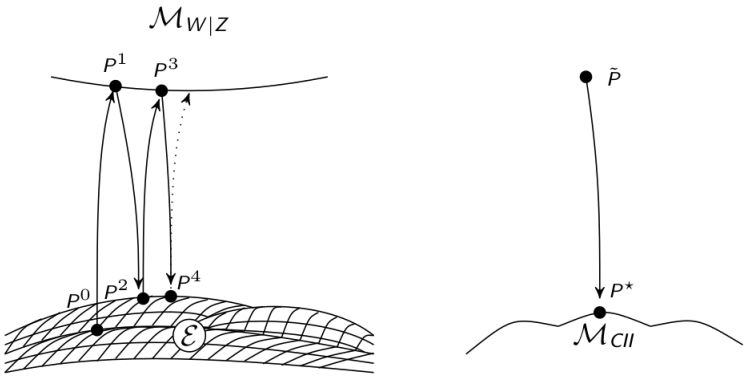
Now we define Q_1 as the m-projection of P_1 to $\mathcal{E}$

$$Q_1 = \arg\min_{Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P_1 \parallel Q).$$

Repeating this leads to

$$P_{i+1} = \arg\min_{P\in\mathcal{M}_{W|Z}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q_i), \quad Q_{i+1} = \arg\min_{Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P_{i+1} \parallel Q).$$

The correspondence between these projections in the bigger space $\mathcal{P}(\mathcal{Z}\times\mathcal{W})$ and one m-projection in $\mathcal{P}(\mathcal{Z})$ is illustrated in Figure 12.



$$\min_{P\in\mathcal{M}_{W|Z},Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) = \min_{Q\in\mathcal{M}_{CII}} D_{\mathcal{Z}}(\tilde{P} \parallel Q)$$

Figure 12: em-Algorithm

The algorithm iterates between the bigger spaces $\mathcal{M}_{W|Z}$ and $\mathcal{E}$ on the left of Figure 12. Using Theorem 2.0.3 we gain that this minimization is equivalent to the minimization between $\tilde{P}$ and $\mathcal{M}_{CII}$. The convergence of this algorithm is given by the following result.

Proposition 3 (Theorem 8 from [3])**.** *The monotonic relations*

$$D_{\mathcal{Z}\times\mathcal{W}}(P_i \parallel Q_i) \geqslant D_{\mathcal{Z}\times\mathcal{W}}(P_{i+1} \parallel Q_i) \geqslant D_{\mathcal{Z}\times\mathcal{W}}(P_{i+1} \parallel Q_{i+1})$$

hold, where equality holds only for the fixed points $(\hat{P},\hat{Q}) \in \mathcal{M}_{W|Z} \times \mathcal{E}$ of the projections

$$\hat{P} = \operatorname*{arg\,min}_{P\in\mathcal{M}_{W|Z}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel \hat{Q})$$
$$\hat{Q} = \operatorname*{arg\,min}_{Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(\hat{P} \parallel Q).$$

Proof of Proposition 3. This is immediate, because of the definitions of the e- and m-projections. □

Hence this algorithm is guaranteed to converge towards a minimum, but this minimum might be local. We will see examples of that in Section 2.1.2.

In order to use this algorithm to calculate Φ_{CII} we first need to determine how to perform an e- and m-projection in this case. The e-projection from $Q \in \mathcal{E}$ to $\mathcal{M}_{W|Z}$ is given by

$$P(z,w) = \tilde{P}(z)Q(w|z),$$

for all $(z,w) \in \mathcal{Z}\times\mathcal{W}$. This is the projection because of the following equality

$$\begin{aligned} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{P(z,w)}{Q(z,w)} \\ &= \sum_{z\in\mathcal{Z}} \tilde{P}(z)\, log\, \frac{\tilde{P}(z)}{Q(z)} + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{P(w|z)}{Q(w|z)}. \end{aligned}$$

The first addend is a constant for a fixed distribution $\tilde{P}$ and the second addend is equal to 0 if and only if $P(w|z) = Q(w|z)$. Note that this means that the conditional expectation of W remains fixed during the

e-projection. This is an important point, because this guarantees the equivalence to the EM algorithm and therefore the convergence towards the MLE. For a proof and examples see Theorem 8.1 in [1] and Section 6 in [2].

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

Proposition 4. *The m-projection from $P \in \mathcal{M}_{W|Z}$ is given by*

$$Q(z,w) = P(x)\prod_{i=1}^{n} P(y_i|x_i,w)P(w)$$

for all $(z,w) \in \mathcal{Z} \times \mathcal{W}$.

The last remaining decision to be made before calculating Φ_{CII} is the choice of the initial distribution. Since it depends on the initial distribution whether the algorithm converges towards a local or global minimum, it is important to take the minimal outcome of multiple runs. One class of starting points that immediately lead to an equilibrium which is in general not minimal are the ones in which Z and W are independent $P^0(z,w) = P^0(z)P^0(w)$. It is easy to check that the algorithm converges here to the fixed point $\hat{P}$

$$\hat{P}(z,w) = \tilde{P}(x)\frac{1}{|\mathcal{W}|}\prod_{i}^{n}\tilde{P}(y_i|x_i)$$
$$\hat{P}(z) = \tilde{P}(x)\prod_{i}^{n}\tilde{P}(y_i|x_i).$$

Note that this is the result of the m-projection of $\tilde{P}$ to $\mathcal{M}_{SI}$, the manifold belonging to Φ_{SI}.

2.1 Comparison

In order to compare the different measures, we need a setting in which we generate the probability distributions of full systems. We chose to use weighted Ising models as described in the next section.

2.1.1 Ising model

The distributions used to compare the different measures in the next chapter are generated by weighted Ising models, also known as binary auto-logistic models as described in [28] Example 3.2.3. Let us consider n binary variables $X = (X_1, \dots, X_n)$, $\mathcal{X} = \{-1, 1\}^n$. The matrix $V \in \mathbb{R}^{n\times n}$ contains the weights v_{ij} of the connection from X_i to Y_j as displayed in Figure 13. Note that this figure is not a graphical model corresponding to the stationary distribution, but merely displays the connections of the conditional distribution of $Y_i = y_i$ given $X = x$ with the respective weights

$$P(y_j|x) = \frac{1}{1 + e^{-2\beta \sum_{i=1}^{n} v_{ij}x_iy_j}}. \tag{9}$$

The inverse temperature $\beta > 0$ regulates the coupling strength between the nodes. For β close to zero the different nodes are almost independent and as β grows the connections become stronger.

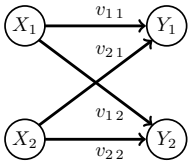


Figure 13: The weights corresponding to the connections for $n = 2$.

We are calculating the stationary distribution $\hat{P}$ by starting with a random initial distribution P^0 and then multiplying by (9) in the following way

$$P^{t+1}(x) = \sum_{x \in \mathcal{X}} P^t(x) \cdot \prod_{j=1}^{n} P(y_i|x).$$

This leads to

$$\hat{P} = \lim_{t \to \infty} P^t.$$

There always exists a unique stationary distribution, see for instance [28], Theorem 5.1.2 .

2.1.2 Results

In this section we are going to compare the different measures experimentally. The code is available at [17]. To distinguish between the Causal Information Integration Φ_{CII} calculated with different sized state spaces of W, we will denote

$$\Phi^m_{CII} = \min_{Q \in \mathcal{M}^m_{CII}} D_{\tilde{Z}}(\tilde{P} \parallel Q).$$

We start with the smallest example possible, with $n = 2$, and the weight matrix

$$V = \begin{pmatrix} 0.0084181 & -0.2401545 \\ 0.39270161 & 0.37198751 \end{pmatrix}$$

shown in Figure 14. In this example every measure is bounded by Φ_I and the measures Φ_I, Φ_G and Φ_{SI} display a limit behavior different from Φ_{CIS} and the Φ_{CII}. The state spaces of W have the size 2, 3, 4, 36 and 92 and the respective measures are displayed in shades of blue that get darker as the state space gets larger. In every case the em-algorithm has been run 10 times with a random input distribution in order to find a global minimum. On the right side of this figure, we are able to see the difference between Φ_{CIS} and Φ_{CII}. Considering the precision of the algorithms we assume that a difference smaller than 5e-07 is approx. zero. We can see that in a region from $\beta = 4$ to $\beta = 6$ the measures differ even in the case of 92 hidden states. So this small case already hints towards $\mathcal{M}_{CII} \subsetneq \mathcal{M}_{CIS}$.

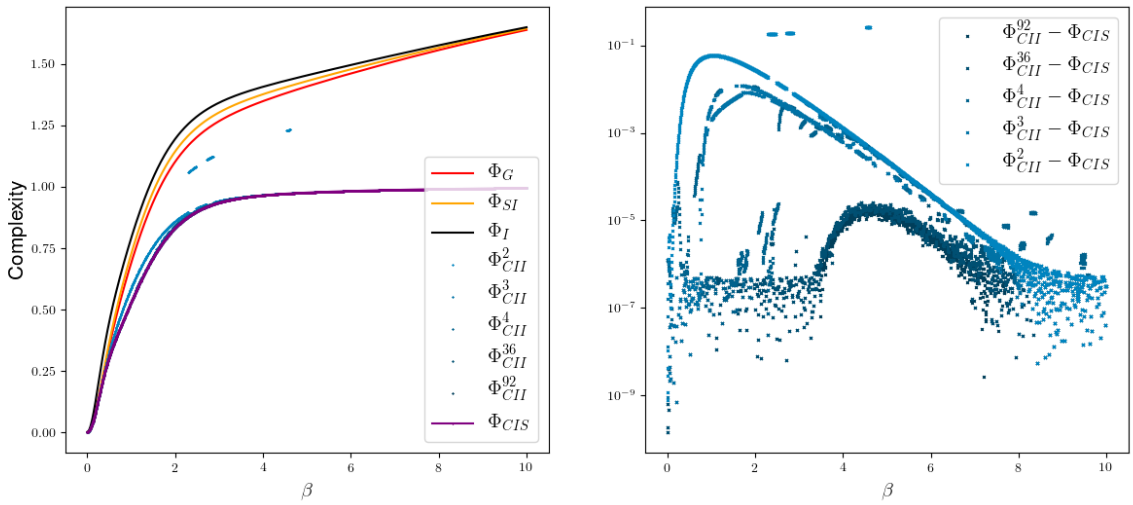


Figure 14: Ising model with 2 nodes and the differences between Φ_{CIS} and Φ_{CII}

Increasing n to 3 makes the difference even more visible, as we can see in Figure 15 produced with the weight matrix

$$V = \begin{pmatrix} -0.43478388 & 0.47448218 & 0.36808313 \\ 0.52117467 & 0.00672578 & -0.7387737 \\ -0.56114795 & -0.96941243 & -0.76408711 \end{pmatrix}.$$

Here we are able to see a difference in the behavior of Φ_G compared to the other measures, since we see that Φ_I, Φ_{SI}, Φ_{CII} and Φ_G are still increasing around $\beta \approx 1.1$, while Φ_G starts to decrease.

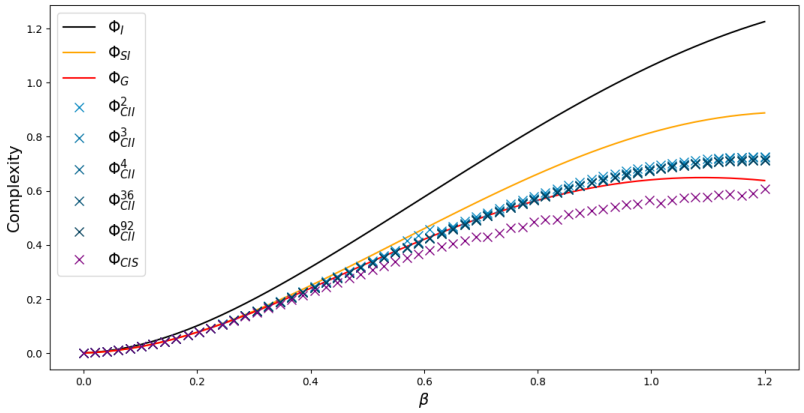


Figure 15: Ising model with 3 nodes.

Now, we are going to focus on an example with 5 nodes. Since it is very time consuming to calculate Φ_{CIS} for more than 3 nodes, we are going to restrict attention to Φ_I, Φ_G, Φ_{SI} and Φ_{CII}. The weight

matrix

$$V = \begin{pmatrix} -0.35615839 & -0.09775903 & 0.89743801 & -0.00604247 & -0.03897772 \\ -0.2260056 & 0.47769717 & -0.4302256 & 0.18692707 & 0.25140741 \\ -0.86081159 & -0.18348132 & -0.71528754 & -0.08100602 & -0.64364176 \\ -0.13967234 & -0.03233011 & -0.81057654 & -0.33327558 & -0.57447322 \\ 0.18920264 & -0.99054716 & 0.32088358 & 0.69100397 & -0.69206604 \end{pmatrix}$$

produces the Figure 16. This example shows that Φ_{SI} is not bounded by Φ_I and therefore does not satisfy Property 2.

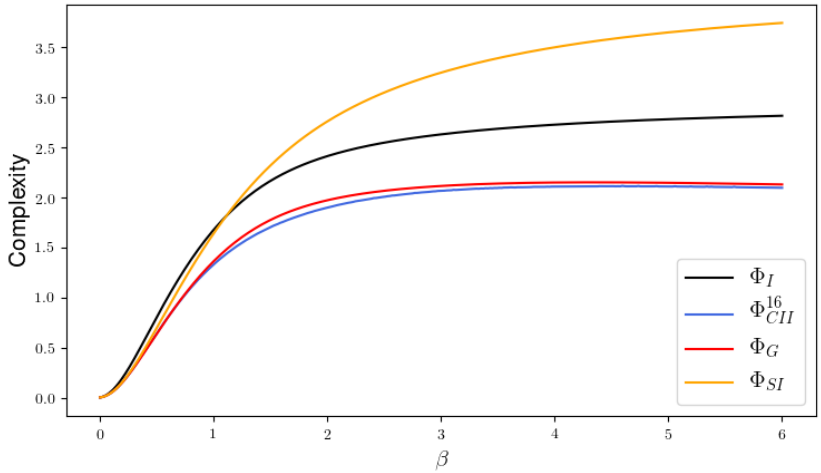


Figure 16: Ising model with 5 nodes.

Using this example, we are going to take a closer look at the local minima the em-algorithm converges to. Considering only Φ_{CII} and varying the size of the state space leads to the upper picture in Figure 17. This figure displays ten different runs of the em-algorithm with each size of state space in different shades of the respective color, namely blue for Φ_{CII}^2, violet for Φ_{CII}^4, red for Φ_{CII}^8 and orange for Φ_{CII}^{16}. We are able to observe how increasing the state space leads to a smaller value of Φ_{CII}. Additionally, the differences between the minimal values corresponding to each state space grow smaller and converge as the state spaces increase.

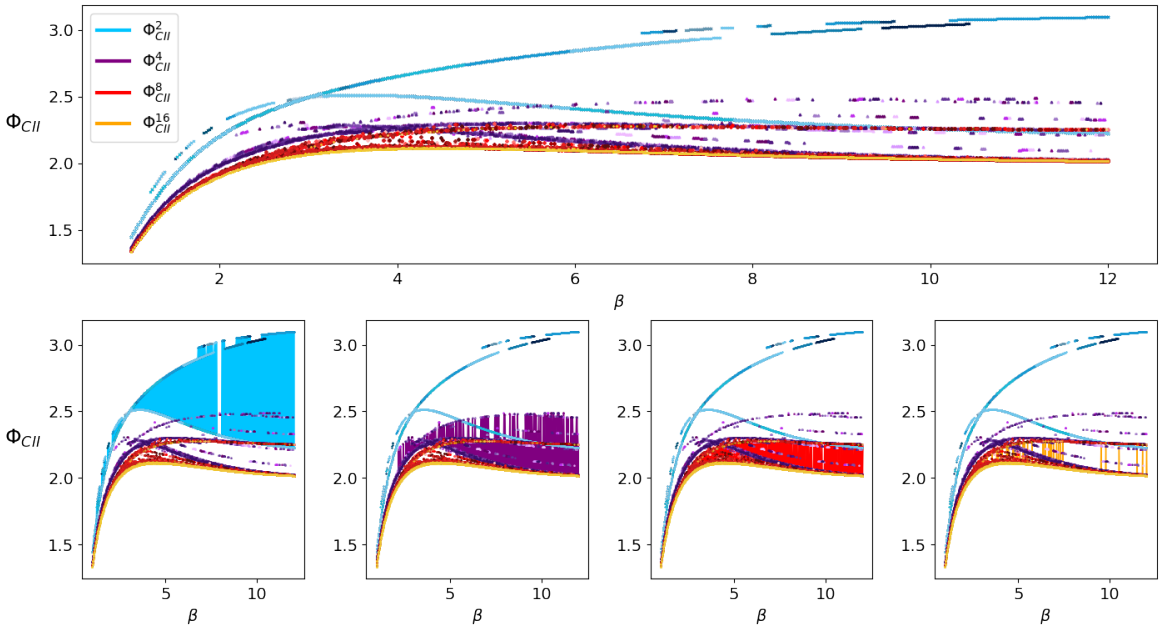


Figure 17: The effect of a different sized state space.

The bottom half of Figure 17 highlights an observation that we made. Each of the four illustrations is a copy of the one above, where the difference between the minima are shaded in the respective color. By increasing the size of the state space the difference in value between the various local minima decreases visibly. We think this is consistent with the general observation made in the context of high dimensional optimization, e. g. [10] in which the authors conjecture that the probability of finding a high valued local minimum decreases when the network size grows.

Letting the algorithm run only once with $|\mathcal{W}| = 2$ on the same data leads to a curve on the left in Figure 18.

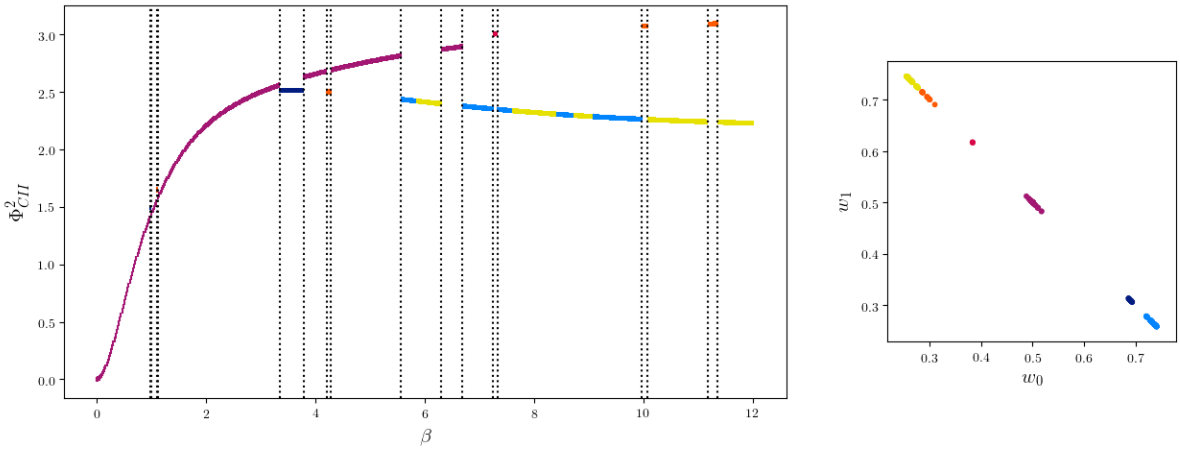


Figure 18: Curve of one run of the em-algorithm for each β coloured according to the distribution of W.

The sets $\mathcal{E}$ defined in (8) and $\mathcal{M}_{CII}$ (5) do not change for different values of β and therefore we have a fixed set of local minima for a fixed state space of W. What does change with different β is which of the local minima are global minima. The vertical dotted lines represent the steps P^t to P^{t+1} in which the KL-divergence between the projection to $\mathcal{M}_{CII}$ is greater than 0.2

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

Meaning that inside the different sections of the curve, the projections to $\mathcal{M}_{CII}$ are close. As β increases, a different region of local minima becomes global. A sketch of this is shown in Figure 19.

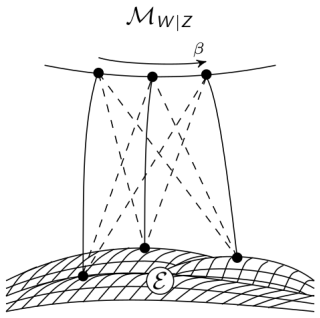


Figure 19: Sketch of different local Minima.

The curve is colored according to the distribution of W as shown on the right side of the figure. We see that a different distribution on $\mathcal{W}$ results in a different minimum, except for the region between 7.5 and 8. The colors light blue and yellow refer to distributions on $\mathcal{W}$ that are different, but symmetric in the following way. Consider two different distributions $Q, \hat{Q}$ on $\mathcal{Z} \times \mathcal{W}$ such that

$$Q(z, w_1) = \hat{Q}(z, w_2) \text{ and } Q(z, w_2) = \hat{Q}(z, w_1)$$

for all $z \in \mathcal{Z}$. Then the corresponding marginalized distributions in $\mathcal{M}^2_{CII}$ are equal

$$\sum_w Q(z, w) = \sum_w \hat{Q}(z, w_1).$$

This symmetry is the reason for the different colors in the region between 7.5 and 8.

Using this geometric algorithm we therefore gain a notion of the local minima on $\mathcal{E}$.

3 Discussion

This article discusses a selection of existing complexity measures in the context of Integrated Information Theory that follow the framework introduced in [5], namely Φ_{SI}, Φ_G and Φ_{CIS}. The main contribution is the proposal of a new measure, Causal Information Integration Φ_{CII}.

In [22] and [4] the authors postulate a Markov condition and an upper bound, given by the mutual information Φ_I, for valid Integrated Information measures. Although Φ_{SI} is not bounded by Φ_I, as we see in Figure 16, it does measure the intrinsic causal connections in a setting in which there exists no common exterior influences. Therefore the authors of [16] criticize this bound. Since wrongly assuming the existence of a common exterior influence might lead to a value that does not measure all the intrinsic causal influences, the question which measure to use strongly depends on the setting we are in. We argue

that using Φ_I as an upper bound in the cases in which we have a common exterior influence is reasonable. The measure Φ_G attempts to extend Φ_{SI} to a setting with exterior influences, but it does not satisfy the Markov condition postulated in [22].

One measure that fulfills all the requirements of this framework is Φ_{CIS}, but it has no graphical representation. Hence the nature of the measured information flow is difficult to analyze. We present in Example 1 a submodel of $\mathcal{M}_{CIS}$, which does not fit into the framework of Integrated Information. For discrete variables Φ_{CIS} does not have a closed form solution and has to be calculated numerically.

We propose a new measure Φ_{CII} which also satisfies all the conditions and has additionally a graphical and intuitive interpretation. Numerically calculated examples indicate that $\Phi_{CII} \subsetneq \Phi_{CIS}$. The definition of Φ_{CII} explicitly includes an interior influence as a latent variable and therefore aims at only measuring intrinsic causal influences. This measure should be used in the setting in which there exists a common exterior influence which is unknown. By assuming the existence of a ground truth, we are able to prove that our new measure is bounded from above by the ultimate value of Integrated Information Φ_T of this system. Although Φ_{CII} also has no analytical solution, we are able to use the em-algorithm to calculate it. The em-algorithm is guaranteed to converge towards a minimum, but this might be local. In our experience the em-algorithm seems to be more reliable and for larger networks faster than the numerical methods we used to calculate Φ_{CIS}. Additionally, by letting the algorithm run multiple times we are able to gain a notion on how the local minima in $\mathcal{E}$ are related to each other as demonstrated in Figure 18 .

Acknowledgement

The authors acknowledge funding by Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

A Graphical Models

Graphical models are a useful tool to visualize conditional independence structures. In this method a graph is used to describe the set of distributions that factor according to it. In our case, we are considering chain graphs.These are graphs, with vertex set V and edge set $E \in V \times V$, consisting of directed and undirected edges such that we are able to partition the vertex set into subsets $V = V_1 \cup \cdots \cup V_m$, called chain components, with the properties that all edges between different subsets are directed, all edges between vertices of the same chain component are undirected and that there are no directed cycles between chain components. For a vertex set τ, we will denote by $pa(\tau)$ the set of parents of element in τ, which are vertices α with a directed arrow from α to an element of τ. Vertices connected by an undirected edge are called neighbours. A more detailed description can be found in [18].

Definition 6. Let T be the set of chain components. A distribution factorizes with respect to a chain graph G if the distribution can be written as follows

$$P(z) = \prod_{\tau \in T} P(x_\tau | x_{pa(\tau)}),$$

where the structure of $P(x_\tau | x_{pa(\tau)})$ can be described in more detail. Let $A(\tau), \tau \in T$ be the set of all subsets of $\tau \cup pa(\tau)$, that are complete in a graph $\tau_\star$, which is an undirected graph with the vertex set $\tau \cup pa(\tau)$ and the edges are the ones between elements in $\tau \cup pa(\tau)$ that exist in G and additionally the ones between elements in $pa(\tau)$. An undirected graph is complete if every pair of distinct vertices is connected by an edge. Then there are non-negative functions ϕ_a such that

$$P(x_\tau | x_{pa(\tau)}) = \prod_{a \in A(\tau)} \phi_a(x).$$

If τ is a singleton then $\tau_\star$ is already complete. There are different kinds of independence statements a chain graph can encode, but we only need the global chain graph markov property. In order to define this property we need the concepts ancestral set and moral graph.

The boundary $bd(A)$ of a set $A \subseteq V$ is the set of vertices in $V \backslash A$ that are parents or neighbours to vertices in A. If $bd(\alpha) \subseteq A$ for all $\alpha \in A$ we call A an ancestral set. For any $A \subseteq V$ there exists a smallest ancestral set containing A, because the intersection of ancestral sets is again an ancestral set. This smallest ancestral set of A is denoted by $An(A)$.

Let G be a chain graph. The moral graph of G is an undirected graph denoted by G^m that consists of the same vertex set as G and in which two vertices α, β are connected if and only if either they were already connected by an edge in G or if there are vertices γ, δ belonging to the same chain component such that $\alpha \to \gamma$ and $\beta \to \delta$.

Definition 7 (Global Chain Graph Markov Property)**.** Let P be a distribution on $\mathcal{Z}$ and G a chain graph. P satisfies the global chain Markov property, with respect to G, if for any triple (Z_A, Z_B, Z_S) of disjoint subsets of Z such that Z_S separates Z_A from Z_B in $(G_{An(Z_A \cup Z_B \cup Z_S)})^m$, the moral graph of the smallest ancestral set containing $Z_A \cup Z_B \cup Z_S$,

$$Z_A \perp\!\!\!\perp Z_B \mid Z_S$$

holds.

Since we are only considering positive discrete distributions, we have the following result.

Lemma 1. *The global chain Markov property and the factorization property are equivalent for positive discrete distributions.*

Proof of Lemma 1. Theorem 4.1 from [15] combined with the Hammersley–Clifford theorem, e.g. Theorem 2.9 in [7], proves this statement. □

In order to understand the conditional independence structure of a chain graph after marginalization, we need the following alogrithm from [23]. This algorithm converts a chain graph with latent variables into a chain mixed graph with the conditional independence structure of the marginalized chain graph. A chain mixed graph has in addition to directed and undirected edges also bidirected edges, called arcs. The condition that there are no semi-directed cycles also applies to chain mixed graphs.

Definition 8. Let M be the set of vertices over which we want to marginalize. The following algorithm produces a chain mixed graph (CMG) with the conditional independence structure of the marginalized chain graph.

1. Generate an ij edge as in Table 2, steps 8 and 9, between i and j on a collider trislide with an endpoint j and an endpoint in M if the edge of the same type does not already exist.
2. Generate an appropriate edge as in Table 2, steps 1 to 7, between the endpoints of every tripath with inner node in M if the edge of the same type does not already exist. Apply this step until no other edge can be generated.
3. Remove all nodes in M.

Conditional independence in CMGs is defined using the concept of c-separation, see for example [23] in Section 4. For this definition we need the concepts of a walk and of a collider section. A walk is a list of vertices $\alpha_0, \ldots, \alpha_k$, $k \in \mathbb{N}$, such there is an edge or arrow from α_i to α_{i+1}, $i \in \{0, \ldots, k-1\}$. A set of vertices connected by undirected edges is called a section. If there exists a walk including a section such that an arrow points at the first and last vertices of the section

$$\to \bullet - \cdots - \bullet \leftarrow$$

then this is called a collider section.

1	i ← m ← j	generates	i ← j
2	i ← m – j	generates	i ← j
3	i ↔ m —j	generates	i ↔ j
4	i ← m → j	generates	i ↔ j
5	i ← m ↔ j	generates	i ↔ j
6	i – m ← j	generates	i ← j
7	i – m – j	generates	i–j
8	m → i – … – ○ ← j	generates	i ← j
9	m → *i* – – ⋯ – –○ ↔ j	generates	i ↔ j

Table 2: Types of edge induced by tripaths with inner node m ∈ M and trislides with endpoint m ∈ M.

Definition 9 (c-separation)**.** Let A, B and C be disjoint sets of vertices of a graph. A walk π is called a c-connecting walk given C, if every collider section of π has a node in C and all non-collider sections are disjoint. The nodes A and B are called c-separated given C if there are no c-connecting walks between them given V and we write $A \perp\!\!\!\perp_c B|C$.

B Proofs

Proof of the Relationship (4). For $n = 2$ this is immediate. Let now $n \geqslant 3$ and $i, j, k \in \{1, \dots, n\}$, $i \neq j \neq k \neq i$. Applying (1) two times leads to

$$\begin{aligned} Q(y_j, x) &= \frac{Q(y_j, x_{I\setminus\{i\}})Q(x)}{Q(x_{I\setminus\{i\}})} \\ Q(y_j, x) &= \frac{Q(y_j, x_{I\setminus\{k\}})Q(x)}{Q(x_{I\setminus\{k\}})} \\ Q(y_j, x_{I\setminus\{i\}})Q(x_{I\setminus\{k\}}) &= Q(y_j, x_{I\setminus\{k\}})Q(x_{I\setminus\{i\}}) \end{aligned}$$

for all $(x, y_j) \in \mathcal{X} \times \mathcal{Y}_j$. Marginalizing over the elements of $\mathcal{X}_k$ yields

$$\begin{aligned} Q(y_j, x_{I\setminus\{i,k\}})Q(x_{I\setminus\{k\}}) &= Q(y_j, x_{I\setminus\{k\}})Q(x_{I\setminus\{i,k\}}) \\ Q(y_j|x_{I\setminus\{i,k\}}) &= Q(y_j|x_{I\setminus\{k\}}). \end{aligned}$$

Using inductively the remaining relations results in (4). □

Proof of Proposition 1. Let $P \in \mathcal{E}^f$ and $Q \in \mathcal{E}$, then the KL-divergence between the two elements is

$$\begin{aligned} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) &= \sum_{z,w} P(z,w)\, log\, \frac{P(x)\prod_i P(y_i|x,w)P(w)}{Q(x)\prod_i Q(y_i|x_i,w)Q(w)} \\ &= \sum_x P(x)\, log\, \frac{P(x)}{Q(x)} + \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i Q(y_i|x_i,w)} + \sum_w P(w)\, log\, \frac{P(w)}{Q(w)} \\ &\geqslant \sum_x P(x)\, log\, \frac{P(x)}{P(x)} + \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i P(y_i|x_i,w)} + \sum_w P(w)\, log\, \frac{P(w)}{P(w)} \\ &= \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i P(y_i|x_i,w)}. \end{aligned}$$

The inequality holds, because in the first and third addend, we are able to apply that the cross entropy is greater or equal to the entropy and in the second addend we use the log-sum inequality in the following way

$$\begin{aligned}
&\sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i Q(y_i|x_i,w)} - \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i P(y_i|x_i,w)} \\
&\quad = \sum_{x,w} P(x)P(w) \sum_y \prod_i P(y_i,|x,w)\, log\, \frac{\prod_i P(y_i|x_i,w)}{\prod_i Q(y_i|x_i,w)} \\
&\quad \geqslant \sum_{x,w} P(x)P(w) \left(\sum_y \prod_i P(y_i,|x,w) \right) log\, \frac{\sum_y \prod_i P(y_i|x_i,w)}{\sum_y \prod_i Q(y_i|x_i,w)} \\
&\quad = 0.
\end{aligned}$$

Therefore the new integrated information measure results in

$$\min_{Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) = \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i P(y_i|x_i,w)}.$$

This can be rewritten to

$$\begin{aligned}
\sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i|x,w)}{\prod_i P(y_i|x_i,w)} &= \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i,x,w)P(x_i,w)}{\prod_i P(y_i,x_i,w)P(x,w)} \\
&= \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i,x_{I\setminus\{i\}}|x_i,w)P(x_i,w)}{\prod_i P(y_i|x_i,w)P(x,w)} \\
&= \sum_{z,w} P(z,w)\, log\, \frac{\prod_i P(y_i,x_{I\setminus\{i\}}|x_i,w)}{\prod_i P(y_i|x_i,w)P(x_{I\setminus\{i\}}|x_i,w)} \\
&= \sum_i I(Y_i; X_{I\setminus\{i\}}|X_i, W).
\end{aligned}$$

□

Proof of Proposition 2. By using the log-sum inequality we get

$$\begin{aligned}
\Phi_{CII} &= \min_{Q\in\mathcal{M}_{CII}} \sum_z P(z) log \frac{\sum_w P(x) \prod_i P(y_i|x,w)P(w)}{\sum_w Q(x) \prod_i Q(y_i|x_i,w)Q(w)} \\
&\leqslant \min_{Q\in\mathcal{M}_{CII}} \sum_w \sum_z P(z,w) log\, \frac{P(x) \prod_i P(y_i|x,w)P(w)}{Q(x) \prod_i Q(y_i|x_i,w)Q(w)} \\
&= \min_{Q\in\mathcal{E}} D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q).
\end{aligned}$$

The fact that every element of $Q \in \mathcal{E}$ corresponds via marginalization to an element in $\mathcal{M}_{CII}$ and every element in $\mathcal{M}_{CII}$ has at least one corresponding element in $Q \in \mathcal{E}$, leads to the equality in the last row. □

Proof of Proposition 4.

$$
\begin{aligned}
D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{P(z,w)}{Q(x)\prod_{i=1}^{n} Q(y_i|x_i,w)Q(w)} \\
&= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, P(z,w) \\
&\quad + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{1}{Q(x)} \\
&\quad + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \sum_{i=1}^{n} P(z,w)\, log\, \frac{1}{Q(y_i|x_i,w)} \\
&\quad + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{1}{Q(w)}
\end{aligned}
$$

The first addend is a constant for P and the others are cross-entropies which are greater or equal to entropy

$$
\begin{aligned}
D_{\mathcal{Z}\times\mathcal{W}}(P \parallel Q) &\geqslant \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, P(z,w) \\
&\quad + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{1}{P(x)} \\
&\quad + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \sum_{i=1}^{n} P(z,w)\, log\, \frac{1}{P(y_i|x_i,w)} \\
&\quad + \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{1}{P(w)} \\
&= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} P(z,w)\, log\, \frac{P(z,w)}{P(x)\prod_{i=1}^{n} P(y_i|x_i,w)P(w)}.
\end{aligned}
$$

Therefore this projection is unique. □

Proof of Theorem 2.2. We need a way to understand the connections in a graph after marginalization. In [23] Sadeghi presents an algorithm that converts a chain graph to a chain mixed graph that represents the markov properties of the original graph after marginalizing, see Definition 8.

Although the actual set of distributions after marginalizing might be more complicated, it is a subset of the distributions factorizing according to the new graph, if the new graph is still a chain graph. This is due to the equivalence of the global chain Markov property and the factorization property in Lemma 1.

At first we will consider the case of two nodes per time step, $n = 2$. We will take a close look at the possible ways a hidden structure could be connected to the left graph in Figure 20. At first we will look at the possible connections between two nodes, depicted on the right in Figure 20. The boxes stand for any kind of subgraph of hidden nodes such that the whole graph is still a chain graph and the two headed dotted arrows stand for a line, or an arrow in any direction. Consider two nodes A and B, then the connections including a box between the nodes can take one of the five following forms

1. they form an undirected path between A and B

2. they can form a directed path from A to B,

3. they can form a directed path form B to A,

4. there exists a collider or

5. A and B have a common exterior influence.

A collider is a node or a set of nodes connected by undirected edges that have an arrow pointing at the set at both ends

$$\rightarrow \bullet \cdots \bullet \leftarrow .$$

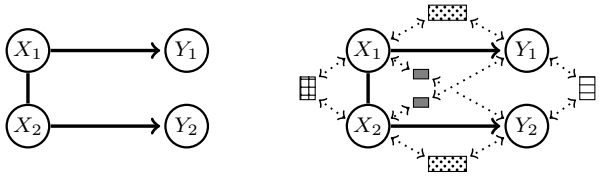


Figure 20: Starting graph and possible two way interactions.

We will start with the gridded hidden structure connected to X_1 and X_2. Since there already is an undirected edge between the X_is an undirected path would make no difference in the marginalized model. The cases (2) and (3) would form a directed cycle which violates the requirements of a chain mixed graph. A collider would also make no difference, since it disappears in the marginalized model. A common exterior influence leads to

$$P(\hat{w})P(x|\hat{w})P(y_1|x_1)P(y_2|x_2) = P(x,\hat{w})P(y_1|x_1)P(y_2|x_2)$$
$$\sum_{\hat{w}} P(x,\hat{w})P(y_1|x_1)P(y_2|x_2) = P(x)P(y_1|x_1)P(y_2|x_2)$$

Now let us discuss these possibilities in the case of a gray hidden structure between X_i and Y_j, $i,j \in \{1,2\}$, $i \neq j$. An undirected edge or a directed edge (3) would create a directed cycle. A directed path (2) from X_i to Y_j would lead to a chain graph in which X_i and Y_j are not conditionally independent given X_j. If there exists a collider (4) in the hidden structure, then nothing else in the graph depends on this part of the structure and it reduces to a factor one when we marginalize over the hidden variables. Therefore the path between X_i and Y_j gets interrupted leaving a potential external influence or effect. Those do not have an additional impact on the marginalized model. A common exterior influence (5) leads to a chain mixed graph which does not satisfy the necessary conditional independence structure, because using the Algorithm 8 leads to an arc between X_i and Y_j, hence they are c-connected in the sense of Definition 9.

The next possibility is a dotted hidden structure between X_i and Y_i, $i \in \{1,2\}$. An undirected path (1) and a directed path (3) would lead to a directed cycle. A directed path (2) would add no new structure to the model since there already is a directed edge between X_i and Y_i. A collider (4) does not have an effect on the marginalized model. Adding a common exterior influence W_1 on X_1, Y_1 results in a new model which is not symmetric in $i \in \{1,2\}$ and does not include $\mathcal{M}_I$, therefore it does not fully contain $\mathcal{M}_{CII}$. By adding additional common exterior W_2 influences on X_2, Y_2 or Y_1, Y_2, in order to include $\mathcal{M}_I$ in the new model, violates the conditional independence statements since nodes in W_1 and W_2 are connected in the moralized graph.

The last hidden structure between two nodes is the striped one between the Y_is. An undirected path (1) or any directed path (2),(3) lead to a graph that does not satisfy the conditional independence

statements. A collider (4) has no impact on the model and a common exterior influence leads to the definition of Causal Information Integration.

Connecting Y_1, Y_2 and $X_i, i \in \{1, 2\}$ leads either to a violation of the conditional independence statements or contains a collider in which case the marginalized model reduces to one of the cases above.

All the possible ways a hidden structure could be connected to three nodes X_1, X_2, Y_1 by directed edges are shown in Figure 21. Replacing any of these edges by an undirected edge would either make no difference or lead to a model that does not satisfy the conditional independence statements. In this case the black boxes represent sections. More complicated hidden structures reduce to this case, since these structures either contain a collider and correspond to one of the cases above or contain longer directed paths in the direction of the edges connecting the structure to the visible nodes, which does not change the marginalized model.

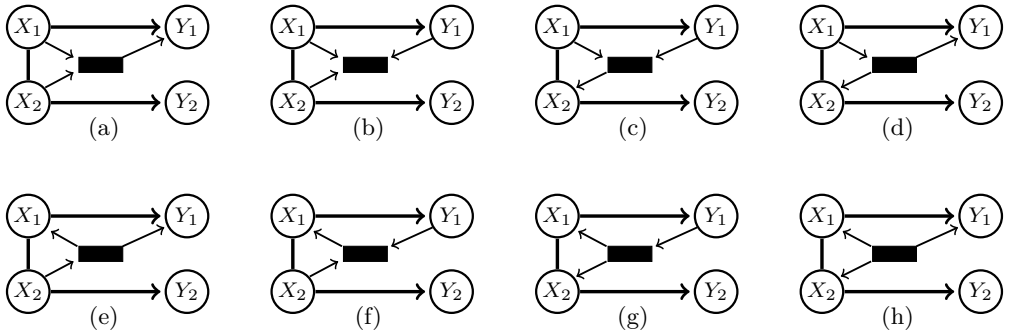


Figure 21: The eight possible hidden structures between three nodes.

The models in (c), (d), (e), (f) and (g) contain either a collider and reduce therefore to one of the cases discussed above or induce a directed cycle. We see that (a) and (h) display structures that do not satisfy the conditional independence statements. The hidden structure in (b) has no impact on the model.

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

Let us now consider a model with $n > 2$. Any hidden structure on this model either connects only up to four nodes and reduces therefore to one of the cases above, contains one of the connections discussed in Figure 21 or only connects nodes among one point in time. The only structures possible to add would be a common exterior influence on the X_is, a common exterior influence on the Y_is or a collider section on any nodes. All these structures do not change the marginalized model. Therefore it is not possible to create a chain graph with hidden nodes in order to get a model strictly larger than $\mathcal{M}_{CII}$. □

References

[1] S. Amari. *Information Geometry and Its Applications*. Springer Japan, 2016.

[2] S. Amari. "Information Geometry of the EM and em Algorithms for Neural Networks". In: *Neural Networks* (1995).

[3] S. Amari, K. Kurata, and H. Nagaoka. "Information geometry of Boltzmann machines". In: *IEEE transactions on neural networks* (1992).

[4] S. Amari, N. Tsuchiya, and M. Oizumi. "Geometry of Information Integration". In: *Information Geometry and Its Applications*. Ed. by N. Ay, P. Gibilisco, and F. Matúš. Springer International Publishing, 2018.

[5] N. Ay. *Information Geometry on Complexity and Stochastic Interaction*. MPI MIS PREPRINT 95, 2001.

[6] N. Ay. "Information Geometry on Complexity and Stochastic Interaction". In: *Entropy* (2015).

[7] N. Ay, J. Jost, H.V. Lê, and L. Schwachhöfer. *Information Geometry*. Springer International Publishing, 2017.

[8] N. Ay, E. Olbrich, and N.A. Bertschinger. "Geometric Approach to Complexity". In: *Chaos (Woodbury, N.Y.)* (2011).

[9] A.B. Barrett and A.K. Seth. "Practical Measures of Integrated Information for Time-Series Data". In: *PLoS Computational Biology* (2011).

[10] A. Choromanska, M. Henaff, M. Mathieu, G.B. Arous, and Y. LeCun. "The Loss Surfaces of Multilayer Networks". In: *Proceedings of Machine Learning Research*. 2015.

[11] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

[12] I. Csiszár and P. Shields. "Information Theory and Statistics: A Tutorial". In: *Foundations and Trends in Communications and Information Theory*. 2004.

[13] I. Csiszár and G. Tusnády. "Information geometry and alternating minimization procedures". In: *Statistics and Decisions* (1984).

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society* (1977).

[15] M. Frydenberg. "The Chain Graph Markov Property". In: *Scandinavian Journal of Statistics* (1990).

[16] M.S. Kanwal, J.A. Grochow, and N. Ay. "Comparing Information-Theoretic Measures of Complexity in Boltzmann Machines". In: *Entropy* (2017).

[17] C. Langer. *Integrated-Information-Measures GitHub Repository*. 2020. URL: `https://github.com/CarlottaLanger/Integrated-Information-Measures`.

[18] S.L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

[19] P.A. Mediano, A.K. Seth, and A.B. Barrett. "Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation". In: *Entropy* (2019).

[20] G. Montúfar. "On the expressive power of discrete mixture models, restricted Boltzmann machines, and deep belief networks - a unified mathematical treatment". PhD thesis. Universität Leipzig, 2012.

[21] M. Oizumi, L. Albantakis, and G. Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0". In: *PLOS Computational Biology* (2014).

[22] M. Oizumi, N. Tsuchiya, S., and Amari. "Unified framework for information integration based on information geometry". In: *PNAS* (2016).

[23] K. Sadeghi. "Marginalization and conditioning for LWF chain graphs". In: *The Annals of Statistics* (2016).

[24] M. Studený. *Probabilistic Conditional Independence Structures*. Springer, 2005.

[25] G. Tononi. "An information integration theory of consciousness". In: *BMC Neuroscience* (2004).

[26] G. Tononi. "Consciousness as Integrated Information: a Provisional Manifesto". In: *Biol. Bull.* (2008).

[27] G. Tononi and G.M. Edelman. "Consciousness and Complexity". In: *Science* (1999).

[28] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer, 2003.