An obstruction theoretic description of integrated information

Torgeir Aambø
Norwegian Defence Research Establishment, FFI

Abstract

We define the integrated cohomology groups of physical systems, and show that integrated information can be detected by the non-vanishing of H^1 obstruction classes that measures the failure of the cause-effect properties of the system to decompose into smaller parts. This allows us to further define higher order integration in a very natural way, and gives a completely functorial notion of integrated information, as suggested by Tull–Kleiner.

1 Introduction

Integrated information theory is a mathematical description of the existing consciousness of a system, developed mainly by Giulio Tononi and his collaborators, CITE . The theory is centered around the computation of Φ, which is a measure of the *integrated information* in the system. Informally, a system S has integrated information, captured by $\Phi(S) > 0$, if the information generated by the entire system's cause-effect structure is greater than the information generated by the combined cause-effect structure of partitions of the system. The formal mathematical framework for computing Φ is rigorous and technical, and uses detailed features of probability theory and measure theory, see CITE

In the present paper we wish to study integrated information through the lens of obstruction theory, which is a mathematical framework studying certain obstruction classes that measures whether a property can be extended through defined levels of sub-structure. They live in mathematical objects called *cohomology groups*, which are abstract formulations of the amount of holes a system has in different dimensions. Obstruction theory was originally developed by Stiefel and Whitney to prove the non-existence of certain vector-fields on manifolds CITE , but has later been applied in a plethora of mathematical contexts CITE .

The conceptual similarity between these two concepts is also extendable to a formal description, which is the aim of this short paper. Building on work by Kleiner and Tull on a categorical description of integrated information theory, see CITE , we assign a simplicial complex $N^\bullet$ to an IIT system S, and use this to build cohomology groups valued in a presheaf that incorporates the cause-effect dynamics of S. Intuitively, we wish to build a mathematical structure whose n-dimensional holes represent n-th order integration, or causal connectedness. We then define a natural obstruction class from any attempted partition ψ of S into subsystems, which we show implies properties for IIT. In particular, if there is a non-trivial obstruction class, then the system has integrated information. These classes will detect whether the global cause-effect dynamics of S can be "glued together" from the local dynamics of its decompo-

sitions. This gluing is described by the presheaf, and the failure to glue is measured by the cohomology.

The upshot of this work is that one can study integrated information from a new lens, and apply fundamental ideas from algebraic topology and homological algebra to study phenomena in consciousness. In particular, our theory gives a much more general approach to integrated information, that side steps a lot of technicalities. We expect that our methods can be combined with topological data analysis, see CITE for an introduction, to locate minimum partitions (MIP) and connect integrated information to new developments in the relationship between topology and information theory, see for example CITE and CITE .

We make no claim surrounding the validity of IIT to explain consciousness and the properties of experience, only to connect some of its mathematical properties to other mathematical concepts known to measure similar features. We view this work not as a new standard for IIT, but as a starting point to explore some of the rich mathematical structures that related to it.

2 IIT systems

In CITE , Tull–Kleiner develop a category theoretic foundation for integrated information theory, based on symmetric monoidal categories with some extra structure that allows for the modeling of cause-effect structures. We will use this as a backdrop for our cohomology theory, hence we quickly review their setup.

2.1 Process theories and decompositions

A *process theory* is a symmetric monoidal category $(\mathcal{C}, \otimes, \mathbb{1})$ together with two designated maps $q_S \colon S \longrightarrow \mathbb{1}$ and $w_S \colon \mathbb{1} \longrightarrow S$, satisfying standard monoidal naturality conditions. Objects in $\mathcal{C}$ are interpreted as system types, and morphisms as physical processes. Morphisms of the form $\mathbb{1} \longrightarrow S$ are called *states*, while morphisms $S \longrightarrow \mathbb{1}$ are called *effects* and morphisms $\mathbb{1} \longrightarrow \mathbb{1}$ are called *scalars*. The designated maps q and w will be referred to as the *discarding effect* and the *noise state* respectively. A morphism $f \colon S \longrightarrow S'$ is *causal* if it preserves the discarding effect,

$$q_{S'} \circ f = q_S,$$

and similarly *cocausal* if it preserves the noise state.

Remark 2.1. Sometimes it is natural to require C to be a symmetric monoidal *dagger* category, which is interpreted as modeling a time-reversal operation.

Example 2.2. The standard example of a process theory is the category of classical probabilistic processes. This is the category Class of finite sets, where a morphism $S \longrightarrow S'$ assigns to an element $s \in S$ an 'unnormalized probability distribution' of elements in S', which are maps $S \times S' \longrightarrow \mathbb{R}^+$. Composition of two maps $f \colon S \times S' \longrightarrow \mathbb{R}^+$ and $g \colon S' \times S'' \longrightarrow \mathbb{R}^+$, is defined by

$$\begin{aligned} f \circ g \colon S \times S'' &\longrightarrow \mathbb{R}^+ \\ (s, s'') &\longmapsto \sum_{s' \in S'} f(s, s') g(s', s''). \end{aligned}$$

The monoidal structure is the cartesian product, meaning the unit $\mathbb{1}$ is the singleton set. Product of morphisms is given by

$$f \otimes g(s, s'')(s', s''') = f(s, s')g(s'', s''').$$

The discarding process q_S is the unique process with $q_S(s) = 1$ for all $s \in S$, meaning a process is causal whenever it is stochastic. The noise state is the uniform probability distribution with $w_S(s) = \frac{1}{|S|}$. This example has a natural time reversal operation, given by $f^\dagger(s, s') = f(s', s)$.

Let $(\mathcal{C}, \otimes, \mathbb{1}, q)$ be a process theory, and S an object in $\mathcal{C}$. A *decomposition* of S is a pair of objects $A, B \in \mathcal{C}$ together with a causal isomorphism $\psi \colon S \simeq A \otimes B$. We denote decompositions by (A, B, ψ), or sometimes simply by (A, B).

Any object S has a decomposition $(S, \mathbb{1})$, which we call the *trivial* decomposition. Two decompositions (A, B, ψ) and (A', B', ψ') are *equivalent* if there are causal isomorphisms $f \colon A \simeq A'$ and $g \colon B \simeq B'$, such that

$$\psi' = (f \otimes g) \circ \psi.$$

The *complement* decomposition of (A, B) is the same decomposition, just in the opposite order; it is denoted $(A, B)^\perp = (B, A)$. If two decompositions are equivalent, then so are their complements.

We define $\mathbb{D}(S)$ to be the set of equivalence classes of decompositions of S. The operation $(-)^\perp$ acts on $\mathbb{D}(S)$, and we define a *decomposition set* of S to be a subset $\mathbb{D} \subseteq \mathbb{D}(S)$ that contains the trivial decomposition and is closed under $(-)^\perp$.

Given a decomposition set $\mathbb{D}$ of S, and a decomposition $(A, B, \psi) \in \mathbb{D}$, we define the *restricted* decomposition set $\mathbb{D}_{|A}$ to be the set of decompositions of A that can be extended to a decomposition of S in $\mathbb{D}$. Furthermore, we define the *restriction* of a state $s \colon \mathbb{1} \longrightarrow S$ to A to be the state defined by

$$s_{|A} = (\mathrm{Id}_A \otimes q_B) \circ \psi \circ s \colon \mathbb{1} \longrightarrow A,$$

and similarly for B. Intuitively, we identify S with its decomposition, and then discard the factor we don't want to consider. By functoriality and naturality, such restrictions depend only on the equivalence class in $\mathbb{D}(S)$, and not on the particular representative.

2.2 System types and directional cuts

To define generalized integrated information theories, we need a notion of systems to study – defined as follows. A *system type* is a triple $(S, \mathbb{D}, T)$, where $S \in \mathcal{C}$, $\mathbb{D}$ is a decomposition set of S and $T \colon S \longrightarrow S$ is a causal process that is interpreted as *time evolution*. The most basic example is the *trivial* system type $(\mathbb{1}, \{\mathbb{1} \otimes \mathbb{1}\}, \mathrm{Id}_\mathbb{1})$.

A state of a system $(S, \mathbb{D}, T)$ is a state of S. Given such a state s and a decomposition $(A, B) \in \mathbb{D}$, CITE constructs a *subsystem* $(A, \mathbb{D}_{|A}, T_{|A})$, where the restricted time evolution is defined by the composition

$$A \xrightarrow{1 \otimes s_{|B}} A \otimes B \xrightarrow{T} A \otimes B \xrightarrow{1 \otimes q_A} A.$$

These subsystems are again independent of the class representative in $\mathbb{D}$.

Given a system type $(S, \mathbb{D}, T)$ and a decomposition $(A, B, \psi) \in \mathbb{D}$, one needs to be able to form systems that disconnects certain causal connections in the total system. IIT 3.0 CITE does this by introducing *cut systems*, that directionally disconnects causal flow from the two parts. IIT 4.0 CITE does this in a slightly different manner, by replacing parts of the inputs of the time evolutions with independent noise. We have included versions of IIT 4.0's disconnections in arbitrary system states for reference and intuition, but we remark that the actual implementation of such disconnections does not matter much, as long as they are functorial.

The B-cut replaces the contribution of B to the time evolution with the noise state, meaning that only the contribution from A matters. It can formally be described as follows:

$$T^{\psi,\leftarrow} \colon A \otimes B \xrightarrow{1 \otimes w_B q_B} A \otimes B \xrightarrow{T} A \otimes B$$

In spirit this forces the time evolution to retain all self-influences from A, and from A to B, but removes all influences from B to A. The arrow in the notation is meant to denote the direction that is removed. Similarly, the A-cut is defined by

$$T^{\psi,\rightarrow} \colon A \otimes B \xrightarrow{w_A q_A \otimes 1} A \otimes B \xrightarrow{T} A \otimes B,$$

and the A-B-cut by

$$T^{\psi,\leftrightarrow} \colon A \otimes B \xrightarrow{1 \otimes w_A \otimes w_B \otimes 1} A \otimes B \otimes A \otimes B \xrightarrow{T \otimes T} A \otimes B \otimes A \otimes B \xrightarrow{1 \otimes q_B \otimes q_A \otimes 1} A \otimes B.$$

The A-B-cut corresponds to Tull–Kleiner's bidirectional cut, where one evolves the system as if A and B were completely disconnected from each other.

We will use the notation $T^{\psi,\delta}$, with $\delta \in \{\leftarrow, \rightarrow, \leftrightarrow\}$ to denote a general disconnection based on a decomposition $\psi \colon S \simeq A \otimes B$. As we did above with the restricted time evolution, we also get definitions of restricted cuts: $T^{\psi,\delta}_{|A}$.

3 Integrated cohomology and obstruction classes

In this section we present the more novel parts of this paper, which is a cohomology theory for IIT systems. We then build an obstruction theory for decomposing a physical system based on this cohomology theory, and show that it detects the existence of integrated information. Lastly we extend this to naturally define higher order integration, often referred to as higher *synergy*.

Remark 3.1. Our setup is formally analogous to Abramsky–Brandenburger's sheaf theoretic description of contextuality, see CITE and CITE , but differs in the added dynamics of the systems. However, we feel that the similarities are enticing, and that the possible interactions are worth further studies.

Given a system type $(S, \mathbb{D}, T)$ and a global state $s \colon \mathbb{1} \longrightarrow S$, we define the category of subsystems of S to be the poset category defined by the poset structure in CITE . In other words, $\mathrm{Sub}(S)$ is the category consisting of subsystems as objects, and morphisms being single comparison maps by the preorder structure.

The idea is then as follows:

1. define a presheaf $\mathcal{F}$ on $\mathrm{Sub}(S)$ that describes the cause-effect dynamics of subsystems of S, where the restriction maps describe subsystem dynamics;
2. define the integrated cohomology of S to be the presheaf-cohomology of the nerve of $\mathrm{Sub}(S)$ valued in $\mathcal{F}$;
3. define natural obstruction classes in $H^1(N^\bullet(\mathrm{Sub}(S)); \mathcal{F})$;
4. prove that these obstruction classes detect the existence of integrated information in S.

There are several possible choices for what presheaf $\mathcal{F}$ to use, depending on how categorical or how system-type specific one wants to be. If one focuses on actual IIT 4.0 constructions, then perhaps a probability based approach using intrinsic information and intrinsic difference would be a good strategy. We have, however, opted for a more categorical and general approach. This is in part to give integrated information more functorial properties, as discussed by Tull–Kleiner in CITE .

As we want to have access to the time evolutions T and $T^{\psi,\delta}$ for subsystems, it is natural to consider the *endomorphisms* as a natural starting position for the presheaf. As we need to be able to understand the *difference* between T and $T^{\psi,\delta}$, we also need access to a subtraction operation. The standard way to do this is to define the presheaf on objects by

$$\begin{aligned} \mathcal{F}\colon \mathrm{Sub}(S)^{\mathrm{op}} &\longrightarrow \mathrm{Ab} \\ A &\longmapsto \mathbb{Z}[\mathrm{End}(A)] \end{aligned}$$

in other words it is the free abelian group on $\mathrm{End}(A)$. Given a subsystem $U \leqslant A$, we define for any endomorphism $f\colon A \longrightarrow A$ a restriction map

$$\rho^A_U(f) = (\mathrm{Id}_U \otimes q_V) \circ f \circ (\mathrm{Id}_U \otimes w_V),$$

where V is the complement of U, i.e. the object such that $A \simeq U \otimes V$, which exists by the definition of a subsystem. We then linearly extended this to $\mathbb{Z}[\mathrm{End}(A)]$, making $\mathcal{F}$ a presheaf.

Remark 3.2. One could also naturally choose to use the actual state $s_{|V}$ instead of the noise state w_V in the definition of the restriction maps. Doing this would mean that the obstruction theory detects context dependent integration, and not the intrinsic integration of the system. We believe that studying what happens upon making this change is an interesting avenue of future research.

Given a system state $(S, \mathbb{D}, T)$ and state $s\colon \mathbb{1} \longrightarrow S$, we define the *integrated cohomology* of $(S, \mathbb{D}, T)$ to be the $\mathcal{F}$-valued presheaf cohomology of the nerve of $\mathrm{Sub}(S)$:

$$H^k_{int}(S) := H^k(N^\bullet(\mathrm{Sub}(S)); \mathcal{F}).$$

Remark 3.3. It is important to note that this is not simply the topological nerve cohomology of $\mathrm{Sub}(S)$, as this will almost always be trivial. The important part here is the use of the presheaf $\mathcal{F}$, which makes this a highly non-trivial cohomology theory.

The intuition is as follows: The nerve of $\mathrm{Sub}(S)$ gives us the geometric construction *where* integrated information can be tested. It consists formally of chains of substates of S of all lengths, with topological structure that understands how these chains are related and

connected. The presheaf gives us *what* is measured on every level – in our case the system dynamics. The cohomology gives us information about which local dynamics that can be glued together to form a global dynamic. If gluing fails, then there is some level of the system where its global dynamics cannot be described in terms of local dynamics, which is the essence of the system having integrated information.

The *decomposition obstruction classes* of S are defined as the cohomology class of the cochains

$$\Delta^{\psi,\delta}(A) = T_{|A} - T_{|A}^{\psi,\delta}$$

for all substates A, decompositions ψ and directions $\delta \in \{\rightarrow, \leftarrow, \leftrightarrow\}$.

Remark 3.4. The reader paying minute attention has perhaps noticed that mechanisms and purviews have yet to be featured in this framework. Our approach to measuring IIT is structurally somewhat different, as instead of starting with mechanisms and purviews and then aggregate local information upwards, we have the global structure as a reference and let mechaniosms and purviews be derived observables. Hence they do not arise as primitive data, but as local test probes for evaluating the obstruction classes. We will, however, need to define and use these to relate our obstruction classes to the integrated information of S, as we will do in the following section.

3.1 Obstruction classes and IIT

The goal of this section is to prove the following theorem.

Theorem 3.5. *Let $(S, \mathbb{D}, T)$ be a system state and $s\colon \mathbb{1} \longrightarrow S$ a state of S. If there is no integrated information in S, then all of the decomposition obstruction classes vanish. In other words, $\Phi(S) = 0 \implies [\Delta^{\psi,\delta}(A)] = 0$ for all A, ψ and δ. In particular, the existence of non-vanishing obstruction classes implies $\Phi(S) > 0$, and hence that the classes $[\Delta^{\psi,\delta}(A)]$ form obstructions to the decomposition of causal structures.*

In order to do so we need to be able to compute and understand Φ. We will use the method of *generalized IIT's* as developed by Tull–Kleiner in CITE . This ensures that our theory works not only for classical IIT, but also quantum IIT and other variants that one would like to study, CITE .

Generalized IIT's requires systems, as we have already studied, but also *experience spaces* and *cause-effect structures*. The details are described in CITE , but we briefly define these for reference.

An *experience space* is a set $\mathbb{E}$ together an *intensity function* $\|-\|\colon \mathbb{E} \longrightarrow \mathbb{R}^+$, a *distance function* $d(-,-)\colon \mathbb{E} \times \mathbb{E} \longrightarrow \mathbb{R}^+$ and a *scalar multiplication* $\mathbb{R}^+ \times \mathbb{E} \longrightarrow \mathbb{E}$ such that

1. $\|re\| = r\|e\|$
2. $r(se) = (rs)e$
3. $1e = e$

for all $e \in \mathbb{E}$ and $r, s \in \mathbb{R}^+$.

A *mechanism* and a *purview* are both chosen subsystems (as defined earlier) of a given system state $(S, \mathbb{D}, T)$, depending on the same state $s\colon \mathbb{1} \longrightarrow S$. Intuitively a mechanism describes the subset of S that we are "letting make a difference", and the purview described the subset of S that is affected.

Given a mechanism-purview pair M, P we denote the combined decomposition set

$$\mathbb{D}_{|M,P} = \mathbb{D}_{|M} \times \mathbb{D}_{|P}.$$

The trivial decomposition is the product of the trivial decompositions of M and P, i.e., $((M, \mathbb{1}), (P, \mathbb{1}))$.

We can now define *cause-effect repertoirs* of a system S, which are defined as an assignment of an experience space $\mathbb{E}(S)$, as well as for each state $s\colon \mathbb{1} \longrightarrow S$ and for each mechanism-purview pair M, P two elements $cau_s(M, P)$ and $eff_s(M, P)$ in $\mathbb{E}(S)$ together with two decomposition maps from

$$cau, eff\colon \mathbb{D}_{|M,P} \longrightarrow \mathbb{E}(S)$$

sending the trivial decomposition to the chosen elements, i.e. $cau((M, \mathbb{1}), (P, \mathbb{1})) = cau_s(M, P)$ and similarly for eff.