Integrated Information Theory (IIT) with Simple Maths

Grégoire Sergeant-Perthuis, Tonglin Yan, Nils Ruet, Kenneth Williford, David Rudrauf

Integrated Information Theory (IIT) with Simple Maths

Grégoire Sergeant-Perthuis[1,*], Tonglin Yan[3], Nils Ruet[3], Kenneth Williford[4], and David Rudrauf[3]

[1] LCQB Sorbonne Université & OURAGAN team, Inria Paris Paris, France

[2] Philosophy and Humanities, University of Texas at Arlington, Arlington, TX, 76019, USA

[3] CIAMS, Université Paris-Saclay, Orsay Cedex, 91405, France; Université d'Orléans, Orléans, 45067, France

[*]Corresponding author: Grégoire Sergeant-Perthuis, gregoireserper@gmail.com

February 2024

Abstract

This article presents a concise mathematical formulation of Integrated Information Theory (IIT), aimed at making the theory more accessible. IIT is one of the most influential theories of consciousness, and from a computational point of view, it can be difficult and time-consuming to find a clear presentation of the technical details of IIT.

Our presentation builds upon the work by Kleiner and Tull [3] that presents IIT in a clear and unified mathematical framework. We propose in this paper a synthesis that highlights the core formalisms of IIT while setting aside the more philosophical aspects, such as the interpretation of IIT's axioms and postulates. The focus is squarely on the mathematical structure of IIT, utilizing basic but central concepts from probability theory, such as Markov kernels and conditional independence, to articulate how IIT formalizes consciousness. The article discusses the 'cutting' of interactions within a system to isolate and evaluate its integrated information, a procedure central to IIT's procedure for quantifying consciousness. By distilling IIT to its mathematical essence, the article aims to foster broader understanding and stimulate further discussion about the theory's potential as a model for consciousness, inviting future explorations into its relationships with other theories and its implications for understanding conscious processes.

1 Introduction

Integrated Information Theory (IIT) is a complex theoretical framework which aim to offer an account of consciousness [7, 1, 5]. In particular, the formalizations of IIT 3.0 and 4.0 may appear opaque and difficult to access for newcomers [1].

We propose a synthetic mathematical formulation of IIT in a brief and accessible format in order to facilitate its assessment and interpretation as a contender for a mathematical theory of consciousness. Our primary focus is on the formal aspects of IIT rather than its motivation and rationale. The only prerequisites are the notions of Markov kernel, conditional expectation, and independence of two random variables.

We build upon the work of Johannes Kleiner and Sean Tull: "The mathematical structure of integrated information theory" (2021) [3] and [1]. Kleiner and Tull have done an excellent job in clarifying its presentation by distinguishing the general features of the theory and its specificities across versions. Their contribution offers a general and unifying mathematical framework for understanding IIT, allowing for a quantitative formulation of the theory. It can still be perceived as quite sophisticated and difficult to process for a less mathematically knowledgeable audience.

To make the theory more accessible and widely disseminated, we aim for a formulation that is less abstract but sufficiently general to capture in a concise way the formalism and concepts at the core of IIT, with the least possible number of concepts, and in a manner that remains valid irrespective of the specifications of IIT across its various versions. We focus on IIT from the standpoint of classical information theory and probability, setting aside quantum frameworks emerging in the literature.

We envision our presentation as an easy and low-cost entry point for accessing the underlying mathematical and computational aspects of IIT in order to foster interest and heuristic discussion about the substance of the theory from computationally driven researchers.

2 Prerequisite concepts from Probability theory

In this section we present the sufficient mathematical concepts needed to construct IIT's Φ. We will define what a probability distribution is, what a Markov kernel (a stochastic map) is, and what conditional independence is.

Definition 1 (Probability measures)**.** *Let X be a finite set, denote by $\mathbb{P}(X)$ the set of probability measures on X,*

$$p \in \mathbb{P}(X) \iff \forall x \in X, p(x) \geq 0 \text{ and } \sum_{x \in X} p(x) = 1 \tag{1}$$

Remark In this document, all spaces X will be finite spaces, so considerations about their σ-algebras will be omitted; any finite set will come with its discrete $\sigma-$algebra (see [6] to learn about $\sigma-$algebras).

Definition 2 (Markov kernels (stochastic maps))**.** *A Markov kernel or stochastic map T from X to Y, denoted as $T : X \to \mathbb{P}(Y)$, sends any point $x \in X$ to a probability measure $T_x \in \mathbb{P}(Y)$. For $x \in X$ and $y \in Y$, we will denote $T_x(y)$ as $T(y|x)$.*

Markov kernels are the probabilistic analogs of deterministic maps. They send any point x to a collection of possible values, each of which can arise with a certain probability.

Example: Sensors can introduce noise, as when one focuses one's (possibly fatigued) eyes on a specific location in the hope of finding an object. In such cases, there is always some level of uncertainty introduced by the sensors; this results in a "radius" of uncertainty, e.g. represented by the standard deviation around the expected position of the object. Markov kernels allow us to model not only this particular kind of uncertainty but also more general kinds of uncertainty.

In the remaining of the document, T will be viewed as a stochastic evolution, a stochastic dynamic, i.e., the evolution of a system for which there is some degree of uncertainty.

Let S be a finite set, representing a collection of indices, with each index corresponding to a random variable; each $i \in S$ is associated with a random variable $\mathbf{X}_i$ that takes values in X_i. The collection of random variables $\mathbf{X}_S := (\mathbf{X}_i, i \in a)$ takes values in $X_S := \prod_{i \in S} X_i$. For any subset $a \subseteq S$, we will denote by $x_a := (x_i, i \in a)$ the joint configurations of variables $X_i, i \in a$ and similarly by $X_a := \prod_{i \in S} X_i$ the space of all their possible configurations. In this document, random variables will be represented in bold font, while the sets they take their values from will be written in regular (non-bold) font.

For any $a \subseteq S$, its complement in S will be denoted as $\overline{a}$.

Definition 3 (Conditional expectation)**.** *Let $Y = \prod_{i \in S_1} Y_i$ be a finite set and let $p \in \mathbb{P}(Y)$. For any $a \subseteq S_1$, and any function $f : Y \to \mathbb{R}$, one defines the conditional expectation with respect to $\boldsymbol{Y}_a$ as;*

$$\forall y_a \in Y_a, \quad \mathbb{E}[f|\, \boldsymbol{Y}_a](y_a) = \sum_{y_{\overline{a}} \in Y_{\overline{a}}} \frac{f(y_{\overline{a}}, y_a) p(y_{\overline{a}}, y_a)}{\sum_{y_{\overline{a}} \in Y_{\overline{a}}} p(y_{\overline{a}}, y_a)} \tag{2}$$

The conditional expectation captures the amount of randomness left in a collection of variables $\mathbf{Y}_i, i \in S_1$ when some of the variables, $\mathbf{Y}_i, i \in a$, are *known* to take a certain value, e.g., $\mathbf{Y}_a = y_a$.

Proposition 1. *For any probability measures $Q \in \mathbb{P}(X_S)$, and any Markov kernel $T : X_S \to \mathbb{P}(X_S)$, one can define a joint distribution over $X_S \times X_S$ as follows,*

$$\forall x'_S, x_S \in X_S, \quad p(x'_S, x_S) := T(x'_S|x_S) Q(x_S) \tag{3}$$

Proof. This proof is elementary but we state it for the sake of completeness. We must show that summing over x'_S and x_S gives 1. One will note that

$$\sum_{x'_S, x_S} p(x'_S, x_S) = \sum_{x_S} Q(x_S) \sum_{x'_S} T(x'_S | x_S) \tag{4}$$

Now, $\sum_{x'_S} T(x'_S|x_S) = 1$ by definition of T's being a Markov kernel. Hence,

$$\sum_{x'_S, x_S} p(x'_S, x_S) = \sum_{x_S} Q(x_S) = 1 \tag{5}$$

Which ends the proof. □

There are two canonical projections on $X \times Y$: the first one $X \times Y \to X$ that sends (x, y) to x, and the second one $X \times Y \to Y$ that sends (x, y) to y. Let us denote the first projection of $X_S \times X_S \to X_S$ as $X_S^{(1)}$ and the second as $X_S^{(2)}$.

Definition 4 (Building a Markov kernel $X_a \to \mathbb{P}(X_b)$ from a prior Q)**.** *Let $T = X_S \to \mathbb{P}(X_S)$ be a Markov kernel. For any $a \subseteq S$ and $b \subseteq S$, a choice of $Q \in \mathbb{P}(X)$ allows us to derive from T the kernel denoted $T^{Q,a,b} : X_a \to \mathbb{P}(X_b)$, which encodes the effect of the variables X_a on X_b. It is defined as,*

$$\forall x_b^{(2)} \in X_b, \forall x_a^{(1)} \in X_a \quad T^{Q,a,b}(x_b^{(2)} | x_a^{(1)}) := \mathbb{E}[\boldsymbol{X}_b^{(2)} = x_b^{(2)} | \boldsymbol{X}_a^{(1)} = x_a^{(1)}] \tag{6}$$

Its explicit expression is,

$$T^{Q,a,b}(x_b^{(2)} | x_a^{(1)}) := 1/C \sum_{\substack{y_{\overline{b}}^{(2)} \in X_{\overline{b}} \\ y_{\overline{a}}^{(1)} \in X_{\overline{a}}}} T(x_b^{(2)}, y_{\overline{b}}^{(2)} | x_a^{(1)}, y_{\overline{a}}^{(1)}) Q(x_a^{(1)}, y_{\overline{a}}^{(1)}) \tag{7}$$

with the normalizing constant:

$$C = \sum_{\substack{y_{\overline{b}}^{(2)} \in X_{\overline{b}} \\ y_{\overline{a}}^{(1)} \in X_{\overline{a}} \\ x_b^{(2)} \in X_b}} T(x_b^{(2)}, y_{\overline{b}}^{(2)} | x_a^{(1)}, y_{\overline{a}}^{(1)}) Q(x_a^{(1)}, y_{\overline{a}}^{(1)}) \tag{8}$$

In the previous Equation 8, the sums over $y_{\overline{b}}^{(2)}, x_b^{(2)}$ give the value 1, and one can show that $C = \sum_{y_{\overline{a}}^{(1)} \in X_a} Q(x_a^{(1)}, y_{\overline{a}}^{(1)})$

The 'cut' cuts the dynamic of T into two pieces. A graphical representation of such a cut operation is given in Figure 1.

Important remark: In IIT, the a priori distibution Q for defining the kernels $T^{Q,a,b} : X_a \to \mathbb{P}(X_b)$ is the uniform distribution over X_S; in other words,

$$\forall x_S \in X_S, \quad Q(x) = \frac{1}{|X_S|} \tag{9}$$

Example: Let $S = \{1, 2\}$, $X = \{0, 1\}$ and $a = \{1\}$ and $b = \{2\}$. In this case $X_S = \{0, 1\}^2$. Let Q be the uniform distribution over X_S, i.e.,

$$Q(0, 0) = Q(1, 0) = Q(0, 1) = Q(1, 1) = \frac{1}{4} \tag{10}$$

Then,

$$T^{Q,a,b}(x_2|x_1) := \frac{\sum_{y_1,y_2\in\{0,1\}} T(y_1, x_2|x_1, y_2)}{2} \tag{11}$$

3 'Cutting' interactions: the central operation of IIT

In this section, we will use probability kernels between subsets of the variable set S to isolate interactions among variables. This approach allows us to contrast interactions induced by independent local subsets with those induced by the whole set. For a graphical representation see Figure 1.

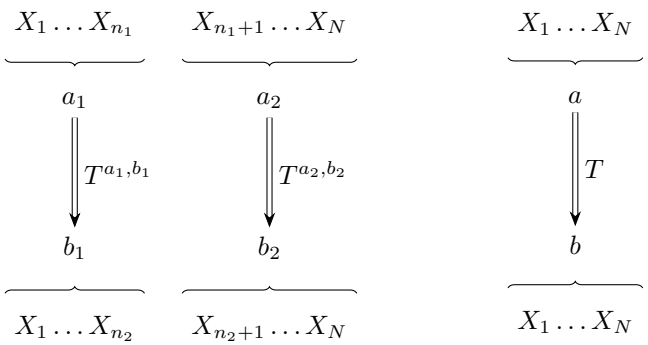

Figure 1: Left: Cutting the interactions. Right: Overall interaction.

Consider two partitions of S: $S = a \cup \overline{a}$ and $S = b \cup \overline{b}$. The dynamic T induces on a and b the dynamic $T^{a,b} : X_a \to \mathbb{P}(X_b)$ and on $\overline{a}$ and $\overline{b}$ the dynamic $T^{\overline{a},\overline{b}}) : X_{\overline{a}} \to \mathbb{P}(X_{\overline{b}})$. We want to build from these two partial dynamics a dynamic on S from $X_S \to \mathbb{P}(X_S)$ that cuts the influence of a on $\overline{b}$ from the one of b on $\overline{a}$. In order to do so we need to create from $(T^{a,b}, T^{\overline{a},\overline{b}})$ a probability kernel from $X_S \to \mathbb{P}(X_S)$; this is done in the next definition.

Definition 5 (Product of local kernels)**.** *For any two probability kernels,* $T^{a,b} : X_a \to \mathbb{P}(X_b)$ *and* $T^{\overline{a},\overline{b}} : X_{\overline{a}} \to \mathbb{P}(X_{\overline{b}})$ *posit,*

$$(T^{a,b} \otimes T^{\overline{a},\overline{b}})(x'|x) := T^{a,b}(x'_b|x_a).T^{\overline{a},\overline{b}}(x'_{\overline{b}}|x_{\overline{a}}) \tag{12}$$

Note that for $x \in X_S$, $T^{a,b} \otimes T^{\overline{a},\overline{b}}(.|x) := T^{a,b}_{x_a} \otimes T^{\overline{a},\overline{b}}_{x_{\overline{a}}}$, where in the right hand $\otimes$ represents the product of two measures (independence).

By extension for any two kernels $T_1 : X \to \mathbb{P}(Y_1)$ and $T_2 : X \to \mathbb{P}(Y_2)$, we define $T_1 \otimes T_2 : X \to \mathbb{P}(Y_1 \times Y_2)$ as $T_1 \otimes T_2(y_1, y_2|x) := T_1(y_1|x)T_2(y_2|x)$.

To compare T and $T^{a,b} \otimes T^{\overline{a},\overline{b}}$ we introduce a 'divergence' on $\mathbb{P}(X_S)$ that allows us to compare distributions. This is the point of the next definition.

Definition 6 (Informal definition of divergence)**.** *For a finite space* Y*, we define a divergence* D *on* $\mathbb{P}(Y)$ *as a function* $D : \mathbb{P}(Y) \times \mathbb{P}(Y) \to \mathbb{R}_{\geq 0}$ *such that, for any two probability distributions* P *and* P_1 *in* $\mathbb{P}(Y)$*,* $D(P, P_1)$ *decreases as the two distributions* P *and* P_1 *get 'closer'; and it reaches its minimum value of* 0 *when and only when* $P = P_1$*.*

Example: any distance on the space of probability measures would make a good divergence (e.g., Wasserstein distance, see [8]). Another candidate for a divergence is the Kullback-Leibler divergence (see [2]), also known as the relative entropy.

For any $x \in X_S$ we can compare T_x and $T^{a,b}_{x_a} \otimes T^{\overline{a},\overline{b}}_{x_{\overline{a}}}$ by computing, $D(T_x|T^{a,b}_{x_a} \otimes T^{\overline{a},\overline{b}}_{x_{\overline{a}}})$.

4 Presenting Φ, focusing on effects

4.1 Little φ_e

When considering a fixed state, denoted as $x \in X_S$, the extent to which the transformation $T^{a,b}_x$ deviates from $T^{a,b}_{x_a} \otimes T^{\overline{a},\overline{b}}_{x_{\overline{a}}}$ reveals the degree to which the dynamics induced by T cannot be simply derived from the dynamics of its constituent parts (a, b) and $(\overline{a}, \overline{b})$. This measure provides valuable insight into the overall "wholeness" of the dynamic behavior of T with respect to its individual components.

The cut operation has a similar definition for a Markov kernel $T : X \to Y$ with $X = \prod_{i \in S} X_i$ and $Y = \prod_{i \in S_1} Y_i$ that are not necessarily the same spaces; let $a \subseteq S$ and $b \subseteq Y$, and $Q_1 \in \mathbb{P}(X)$ a prior, then for a given a prior $S \in \mathbb{P}(Y)$, one defines,

$$T^{Q_1,a,b}(x_b|y_a) := 1/C \sum_{\substack{y_{\overline{a}} \in Y_{\overline{a}} \\ x_{\overline{b}} \in X_{\overline{b}}}} T(x_b, x_{\overline{b}}|y_a, y_{\overline{a}})Q_1(x_a, y_{\overline{a}}) \tag{13}$$

with the normalizing constant:

$$C = \sum_{\substack{y_{\overline{a}} \in Y_{\overline{a}} \\ x_{\overline{b}} \in X_{\overline{b}} \\ x_b \in X_b}} T(x_b, x_{\overline{b}} | y_a, y_{\overline{a}}) Q_1(x_a, y_{\overline{a}}) \tag{14}$$

In what follows the 'cut' operations are applied to the Markov $T^{Q,P,M}$ with $M \subseteq S$ and $P \subseteq S$ of S where Q is the uniform prior on the configurations of X_S. We give a specific notation to this Markov kernel:

$$\forall x_M \in X_M \quad T^P_{M,x_M} := T^{P,M}_{x_M} \tag{15}$$

Then one computes for $a \subseteq M$ and $b \subseteq P$, and $Q_1 \in \mathbb{P}(X_M)$ the uniform distribution on $\mathbb{P}(X_M)$, the associated 'cut' kernel $T_M^{P\,Q_1,a,b} : X_a \to X_b$. We will denote $T_M^{P\,Q_1,a,b}$ as $T_M^{P,(a,b)}$ to recall that the 'cut' is made on the kernel T^P_M, with $a \subseteq M$, $b \subseteq P$.

For simplicity of the presentation, we focus our presentation on the 'effect' component of IIT, the "φ_e" of IIT. We will denote it as φ. From there we will compute the associated Φ. The 'cause' component of IIT follows similar constructions.

We choose to first focus on "φ_e" in this section as it is sufficient to understand the computation of Φ and the motivation for the expression of Φ without getting lost in computational details that are not essential for understanding how Φ is computed.

Definition 7. *For any $M, P \subseteq S$ and $x_M \in X_M$,*

$$\varphi^P_{M,x_M} := \inf_{\substack{a \subsetneq M \\ b \subseteq P}} D(T^P_{M,x_M} | T^{P,(a,b)}_{M,x_a} \otimes T^{P,(\overline{a},\overline{b})}_{M,x_{\overline{a}}}) \tag{16}$$

And

$$\varphi^*_{M,x_M} := \max_{P \subseteq S} \varphi^P_{M,x_M} \tag{17}$$

Notation: Let us denote $\psi(M,x) := \operatorname{argmax} \varphi^P_M$

4.2 Big Φ_e, focusing on effects

See any $M \subseteq S$ as its collection of parts $\mathcal{P}(M) := \{a \subseteq M\}$. Consider $x \in X_S$. Assume that $b \subseteq \psi(M, x_M)$; we will denote $b_a := b \cap \psi(a,x)$ and $\overline{b} \cap \psi(a,x)$ as $\overline{b}_a$ as. Then,

$$\Phi_{M,x,b} := \sum_{a \subseteq M} \varphi^*_{a,x_a} D(T^{\psi(a,x_a)}_{a,x_a} | T^{b_a}_{a,x_a} \otimes T^{\overline{b}_a}_{a,x_a}) \tag{18}$$

It is important to remark that when ϕ_a^* cancels it does not contribute to the previous sum (see comments in Appendix A.4 [3] on that subject).

Then,

$$\Phi_{M,x} = \text{argmin}_{b \subseteq \psi(M,x_M)} \Phi_{M,x,b} \tag{19}$$

and then,

$$\Phi_x := \max_M \Phi_{M,x} \tag{20}$$

5 Discussion

We proposed a concise mathematical formulation of the core mathematical structure of IIT as a tool to facilitate its analysis.

The quantities ϕ, Φ discussed in this article could in principle be computed for any dynamical system, but the associated quantities could be equal to 0.

Although this is beyond the scope of this article, it would be worth for future contributions to further situate, formally and systematically, IIT within the constellation of other theoretical proposals directly or indirectly relevant to consciousness. For instance, recent discussions and non-formal studies have emerged regarding the relationships between IIT and active inference or the Free Energy Principle [4]. Let us just envision here some preliminary considerations in that direction.

For instance, one may wonder how agents that act according to active inference may be imbued with, and leverage an IIT structure.

Let us recall the basic formulation of active inference. There are two parts to active inference: the first one is to compute the prior, and the second one is to optimize preferences through action.

One possible way of fitting IIT's algorithm into such a framework could be to assume that the internal world model of the agent is cut into two pieces, one for external-world dynamics and one for its own internal dynamics. $X = X_{int} \times X_{ext}$, where X_{int} is a collection of variables that account for the internal state of the agent and X_{ext} for the external state. The actions of the agent and the possible evolution of the external world then would induce a dynamics $T_{\theta,a} : X_{int} \times X_{ext} \to \mathbb{P}(X_{int} \times X_{ext})$ that depends on the model θ the agent has of its environment, with $a \in \mathcal{A}$ a choice of action. Then we can consider the effects of the dynamical system T on $X_{int} \to \mathbb{P}(X_{int})$ as discussed in the previous sections, i.e., considering $a = \{X_{int}\} \subseteq \{X_{int}\} \cup \{X_{ext}\}$ and $b = \{X_{int}\}$; we denote the associated dynamics $T^{\text{int}}_{\theta,a}$. At each time t of the algorithm of active inference, after updating the prior θ and choosing an action a, one could then use IIT's computations on $T^{\text{int}}_{\theta,a}$, which would account for the agent's own level of consciousness. The agent would thus use a level of consciousness as defined by IIT as a quantitative feedback. The agent could then regulate its own functioning based on internal quantities representing, at a meta-level, if not its own consciousness, at least some measure of its own efficiency or flow.

6 Acknowledgment

We would like to thank Gaspard Fougea for the insightful discussions on the Axioms of IIT and for his presentation to the PMMC (Paris Mathematical Models of Consciousness seminar). We would like to thank all the speakers of this seminar for the very interesting discussions.

References

[1] Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M Haun, William Marshall, William GP Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjørn E Juel, Shuntaro Sasai, Keiko Fujii, Isaac David, Jeremiah Hendren, Jonathan P Lang, and Giulio Tononi. Integrated information theory (iit) 4.0: Formulating the properties of phenomenal existence in physical terms, 2022.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[3] Johannes Kleiner and Sean Tull. The mathematical structure of integrated information theory. In *Frontiers in Applied Mathematics and Statistics*, 2020.

[4] Christoffer Lundbak Olesen, Peter Thestrup Waade, Larissa Albantakis, and Christoph Mathys. Phi fluctuates with surprisal: An empirical pre-study for the synthesis of the free energy principle and integrated information theory. *PLOS Computational Biology*, 19(10):e1011346, 2023.

[5] Bjorn Merker, Kenneth Williford, and David Rudrauf. The integrated information theory of consciousness: a case of mistaken identity. *Behavioral and Brain Sciences*, 45, 2022.

[6] Walter Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., USA, 1987.

[7] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461, 2016.

[8] Cédric Villani. Optimal transport: Old and new. 2008.