# Paris Mathematical Models of Consciousness: September 29, 2023: Mathematical Aspects of Integrated Information Theory (IIT)

Grégoire Sergeant-Perthuis

October 2023

**Caution:** These are preliminary notes.

We will provide a succinct presentation of Integrated Information Theory (IIT) based on the recent formulations outlined in references [2] and [1]. Our primary focus will be on the formal aspects of IIT rather than its motivational aspects. Our objective is to create a quick-reference document for gaining a detailed understanding of the algorithm that underlies IIT. To achieve this, we will introduce the concepts of probability kernels and conditional independence.

## 1 Definition: Markov kernel (stochastic map), conditional independence

**Definition 1 (Probability measures)** *Let $X$ be a finite set, denote $\mathbb{P}(X)$ 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 consideration about their sigma algebra will be omitted; any finite set will come with is discrete $\sigma-$algebra.

**Definition 2 (Markov kernels (stochastic map))** *A Markov kernel or stochastic map $T$ from $X$ to $Y$, denoted as $T = X \to Y$, 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 $T(y|x)$.*

$T$ can be viewed as a stochastic evolution, dynamic.

Example: Sensors can introduce noise, such as when you focus 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, resulting in a "radius" of uncertainty (represented by the standard deviation) around the expected position of the object.

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 to a random variable $\mathbf{X}_i$ that takes values in $X_i$. The collection of random variables $\mathbf{X}_S := (\mathbf{X}_i, i \in S)$ takes values in $\prod_{i \in S} X_i$. For any subset $a \subseteq S$, we will denote $x_a := (x_i, i \in a)$ and similarly $X_a := \prod_{i \in S} X_i$. In this document, random variables will be represented in bold font, while the sets they take values in will be written in regular (non-bold) font.

For any $a \subseteq S$, it complementary 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;*

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

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

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

There are two canonical projections on $X \times Y$ the first one that sends $X \times Y \to X$ and the second one$X \times Y \to Y$. 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 (Probability Kernel from $X_a \to X_b$)** *Let $T : X_S \to X_S$ be aprobability kernel. For any $a \subseteq S$ and $b \subseteq S$, a choice of $Q \in \mathbb{P}(X)$, allows to derive from $T$ the kernel that encodes the effect of the variables $X_a$ on $X_b$ as,*

$$\forall x_b^{(2)} \in X_b, x_a^{(1)} \in X_a \quad T^{Q,a,b}(x_b^{(2)}|x_a^{(1)}) := \mathbb{E}[X_b^{(2)} = x_b^{(2)}|X_a^{(1)} = x_a^{(1)}] \tag{4}$$

*Explicitly,*

$$T^{Q,a,b}(x_b^{(2)}|x_a^{(1)}) := 1/C \sum_{\substack{y_{\overline{b}}^{(2)} \in X_{\overline{b}} \\ y_{\overline{a}}^{(1)} \in X_{\overline{a}}}} T(x_b^{(2)}, y_{\overline{b}}^{(2)}|x_a^{(1)}, y_{\overline{a}}^{(1)}) Q(x_a^{(1)}, y_{\overline{a}}^{(1)}) \tag{5}$$

*with*

$$C = \sum_{\substack{y_{\overline{b}}^{(2)} \in X_{\overline{b}} \\ y_{\overline{a}}^{(1)} \in X_{\overline{a}} \\ x_b^{(2)} \in X_b}} T(x_b^{(2)}, y_{\overline{b}}^{(2)} | x_a^{(1)}, y_{\overline{a}}^{(1)}) Q(x_a^{(1)}, y_{\overline{a}}^{(1)}) \tag{6}$$

Remark In fact the constant that normalizes the kernel from $X_a \to X_b$ has a simple expression:

$$C = \sum_{y_b^{(1)} \in X_b} Q(x_a^{(1)}, y_{\overline{a}}^{(1)}) \tag{7}$$

**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.

$$Q(0, 0) = Q(1, 0) = Q(0, 1) = Q(1, 1) = \frac{1}{4} \tag{8}$$

Then,

$$T^{Q,a,b}(x_2|x_1) := \frac{\sum_{y_1, y_2 \in \{0,1\}} T(y_1, x_2|x_1, y_2)}{2} \tag{9}$$

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

$$\forall x_S \in X_S, \quad Q(x) = \frac{1}{|X_S|} \tag{10}$$

We denote the uniform distribution $Q$ as $U(X_s)$.

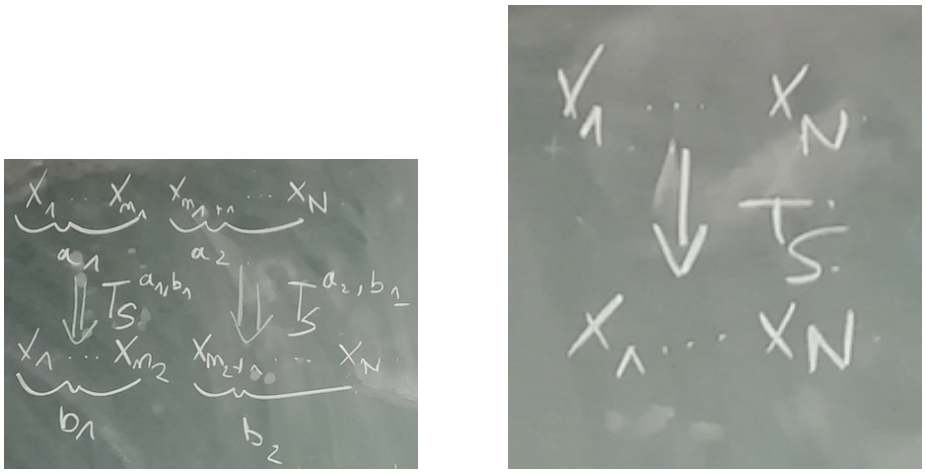
# 2 'Cutting' interactions and wholeness

In this section, we will utilize 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. See Figure 1.

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 $T^{a,b} : X_a \to X_b$ and on $\overline{a}$ and $\overline{b}$ the dynamic $T^{\overline{a},\overline{b}}) : X_{\overline{a}} \to X_{\overline{b}}$. We want to build from these two partial dynamics a dynamic on S from $X_S \to X_S$ 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 $X_S \to X_S$, this is done in the next definition.

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

$$(T^{a,b} \otimes T^{\overline{a},\overline{b}})(x'|x) := T^{a,b}(x'_b|x_a).T^{\overline{a},\overline{b}}(x'_{\overline{b}}|x_{\overline{a}}) \tag{11}$$



(a) Cutting the interactions (b) Whole interaction

Figure 1: Overall Caption for Both Images

Remark 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 $T_1 : X \to Y_1$ and $T_2 : X \to Y_2$ we define $T_1 \otimes T_2 : X \to Y_1 \times Y_2$ 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 propose to introduce a 'divergence' on $\mathbb{P}(X_S)$ that allows to compare distributions.

**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 $P_1$ become 'closer,' and it reaches its minimum value of 0 when and only when $P = P_1$.*

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, $D(T_x | T^{a,b}_{x_a} \otimes T^{\overline{a},\overline{b}}_{x_{\overline{a}}})$.

# 3 Little $\varphi$

When considering a fixed state, denoted as $x \in X_S$, the extent to which the transformation $T^{a,b}_x$ deviates from $T^{a,b}_x \otimes T^{\overline{a},\overline{b}}_x$ 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.

We can extend the concepts we have previously introduced for the subsets $M \subseteq S$ and $P \subseteq S$ of $S$. We pose:

$$\forall x_M \in X_M \quad T^P_{M,x_M} := T^{P,M}_{x_M} \tag{12}$$

For simplicity of presentation, our primary focus in these notes is to elaborate on the 'effect' component, denoted as $\varphi$ (sometimes referred to as $\varphi_e$ in the literature), while we will not delve into the 'cause' $\varphi_c$ in the context of IIT.

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

$$\varphi^P_{M,x_M} := \inf_{\substack{a \subseteq M \\ b \subseteq P}} D(T^P_{M,x_M} | T^{P,(a,b)}_{M,x_a} \otimes T^{P,(\overline{a},\overline{b})}_{M,x_{\overline{a}}}) \tag{13}$$

*And*

$$\varphi^*_{M,x_M} := \max_{P \subseteq S} \varphi^P_{M,x_M} \tag{14}$$

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