Self-conditionally trivial sub-sigma-algebras

Julian Newman

Abstract

In a fairly common "proof" of the Ergodic Decomposition Theorem, for a random variable whose law is preserved under a dynamical system, it is incorrectly assumed that the conditional probability distribution, conditioned on observability of every strictly invariant set's including or not including the value of the random variable, must almost surely agree on the sigma-algebra of strictly invariant sets with the Dirac mass at the true value of the random variable. Although this seemingly natural assumption turns out to be false, the Ergodic Decomposition Theorem nonetheless guarantees the weaker – seemingly unnatural – property that the sigma-algebra of strictly invariant sets must almost surely have a 0-1 law under this conditional probability distribution, even when this 0-1 law disagrees with the Dirac mass at the true value of the random variable. This naturally prompts the question of whether – in greater generality beyond the specific setting of the Ergodic Decomposition Theorem – the property of a sub-sigma-algebra becoming probabilistically trivial under conditioning with respect to itself (without reference to whether this triviality agrees with the Dirac mass at the true sample point) is a "more natural" property than one might first expect. I will present a result in this direction.

Part 1 (Thursday 2nd February 2023)

Setup

- Let X be some (generally, continuous) state space that is "not too horrible" [more precisely: a separable metric space that is a Borel subset of its completion; any space that you might ever want to consider in any practical application will fulfil this].
- Someone picks a point $x \in X$ at random, with probability distribution $\mathbb{P}$.
- I cannot see exactly the chosen point x, but I have some **"limited observability"** of it (which will be defined precisely shortly).
- There is some region of the state space – i.e. a set $B \subset X$ – that is of interest: I am interested in whether the randomly selected point x is in this set B.
- **Question:** What is the **posterior probability** that $x \in B$ given my limited observability?

Standing Assumption. All subsets of X that we mention or consider are *"Borel sets"*.

→ heuristically: they are "not too horrible"; any subset of the state space that you might ever want to consider in practice will fulfil this.

By way of notation: the probability that the randomly selected point lies in a pre-specified set $A \subset X$ is denoted $\mathbb{P}(A)$. (Hence in particular, the "prior probability that $x \in B$" is $\mathbb{P}(B)$.) Furthermore, for any $A \subset X$ with $\mathbb{P}(A) > 0$ (meaning that the probability that $x \in A$ is not infinitesimal), the "probability that $x \in B$ conditional on the observation that $x \in A$" is denoted

$$\mathbb{P}(B|A) \ = \ \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

"Limited observability": this is represented by a collection $\mathcal{C}$ of subsets of X (which we shall call *"test sets"*) where for each member $C \in \mathcal{C}$, I can "observe" or "test" whether or not $x \in C$.

The case that $\mathcal{C}$ is finite

Let us answer the Question in the case that the collection of available test sets is finite, $\mathcal{C} = \{C_1, \ldots, C_n\}$. **We will represent the answer as a function whose input is the invisible true value of x and whose corresponding output is the resulting posterior probability that $x \in B$ given observability of $\mathcal{C}$.** This function will be denoted $\mathbb{P}(B|\mathcal{C})$.

First, we take **the partition of X generated by $\mathcal{C}$**; this is

$$\mathrm{Part}_X(\mathcal{C}) \ := \ \left\{ \bigcap_{i=1}^{n} \chi_{C_i}(a_i) \ : \ (a_i)_{i=1,\ldots,n} \in \{0,1\}^n \right\} \setminus \{\emptyset\}$$

where we use the notation

$$\chi_C(a) \ := \ \begin{cases} C & a = 1 \\ X \setminus C & a = 0. \end{cases}$$

(The reason for the "$\setminus\{\emptyset\}$" at the end is that it is redundant to include the empty set as a member of a partition.)

As an illustrative example (think of the stereotypical "Venn diagram"): Imagine X is a solid rectangle, and $\mathcal{C}$ has three members $C_1, C_2, C_3 \subset X$, which are mutually overlapping discs. If you draw the rectangle and the three overlapping circles, then this pictorially divides the inside of the rectangle into 8 distinct regions; these 8 regions are the members of the partition $\mathrm{Part}_X(\mathcal{C})$.

Definition. Define the set $X_{\mathbb{P},\mathcal{C}} \subset X$ by

$$X_{\mathbb{P},\mathcal{C}} \ := \ \bigcup \{A \in \mathrm{Part}_X(\mathcal{C}) : \mathbb{P}(A) > 0\},$$

that is, $X_{\mathbb{P},\mathcal{C}}$ is the union of all members of $\mathrm{Part}_X(\mathcal{C})$ that have a larger-than-infinitesimal probability of including the randomly selected point; and define the function

$$\mathbb{P}(B|\mathcal{C})\colon X_{\mathbb{P},\mathcal{C}} \to [0,1]$$

such that for every $A \in \mathrm{Part}_X(\mathcal{C})$ with $\mathbb{P}(A) > 0$,

$$\forall\, x \in A, \ \mathbb{P}(B|\mathcal{C})(x) \ = \ \mathbb{P}(B|A).$$

Let us comment on the domain of definition of $\mathbb{P}(B|\mathcal{C})$. If $\mathrm{Part}_X(\mathcal{C})$ includes some sets that are infinitely unlikely for the randomly selected point to land in [e.g. if X is a two-dimensional space with $\mathbb{P}$ a continuous distribution, and there are elements of $\mathcal{C}$ that overlap with each other exactly on some one-dimensional curve], then $\mathbb{P}(B|\mathcal{C})$ is not well-defined at the points that lie in these sets. But, from a practical point of view, this is not a problem, as the likelihood of landing in such a set, i.e. of lying outside the domain of definition of $\mathbb{P}(B|\mathcal{C})$, is infinitesimal.

"Trivial Proposition" 1. *In the above setting where $\mathcal{C}$ is finite, if $B \in \mathcal{C}$ then*

$$\begin{aligned}\mathbb{P}(B|\mathcal{C})(x) &= \begin{cases} 1 & x \in B \\ 0 & x \notin B \end{cases} \\ &= \mathbb{1}_B(x)\end{aligned}$$

for all $x \in X_{\mathbb{P},\mathcal{C}}$.

This "Trivial Proposition" serves as a *sanity check*: if B itself is among the sets for which we can observe whether the randomly selected point lies in the set, then the "posterior probability" that the randomly selected point lies in B should no longer be a matter of "probability" at all but of direct deterministic verification.

What if there are infinitely many test sets?

– **Example:** suppose that

- X is a square, $X = [0,1) \times [0,1)$;
- writing $x = (x_1, x_2)$, we have no observability of the vertical component x_2 of the randomly selected point x, but we *can* observe the horizontal component x_1;

  $\rightarrow$ we have measuring tools that can measure x_1 to arbitrarily high accuracy.
- One way that we can imagine this is as being that the collection of test sets is

$$\mathcal{C} = \{[\tfrac{i}{2^n}, \tfrac{i+1}{2^n}) \times [0,1) : n \geq 1,\, 0 \leq i < 2^n\}.$$

In such a scenario, namely where $\mathcal{C}$ has infinitely many members, can we meaningfully define "the posterior probability that $x \in B$ given the observability of $\mathcal{C}$"?

The answer is, essentially, yes. There are a couple of standard constructions of this, which are equivalent to each other. However, I find these constructions not very intuitive in terms of relating to a physical picture of what is going on. So I have spent some time recently thinking about how to come up with a construction with clearer physical intuition, and the result is what I will now present.

Let $L(\mathbb{P}, [0,1])$ be "the set of all Borel-measurable functions $f\colon X \to [0,1]$ identified up to $\mathbb{P}$-almost-sure equality". Let us clarify a couple of things in this sentence:

- Heuristically, "Borel-measurable" means "not too horrible"; any function on a state space X that you might ever want to consider in practice will fulfil this.

- "identified up to $\mathbb{P}$-almost-sure equality": Suppose we have two Borel-measurable functions $f_1 \colon X \to [0,1]$ and $f_2 \colon X \to [0,1]$; and suppose that

$$\mathbb{P}(\{y \in X : f_1(y) \neq f_2(y)\}) = 0,$$

  i.e. it is infinitely unlikely that the randomly selected point x will have $f_1(x)$ and $f_2(x)$ disagreeing with each other. Then, for all practical intents and purposes, we may regard f_1 and f_2 as being the same function. So then, $L(\mathbb{P}, [0,1])$ is defined such that f_1 and f_2 represent the same element of $L(\mathbb{P}, [0,1])$.

Now we will need to equip $L(\mathbb{P}, [0,1])$ with a topology, i.e. a way of defining what it means for two functions from X to $[0,1]$ to be "very close to each other". The way we will do this is as follows: two elements $f_1, f_2 \in L(\mathbb{P}, [0,1])$ are considered to be "very close to each other" if the following equivalent statements hold:

- for $x \sim \mathbb{P}$, $\mathrm{Prob}(|f_1(x) - f_2(x)| \text{ is very small})$ is very close to 1;
- for $x \sim \mathbb{P}$, $\underbrace{\mathrm{Exp}[|f_1(x) - f_2(x)|^p]}_{=: \mathbb{E}[|f_1 - f_2|^p]}$ is very small (with any fixed $p \in [1, \infty)$).

The former, more precisely, corresponds to the *topology of convergence in probability*, while the latter corresponds to the *topology of L^p-convergence*; but here the two topologies are the same as each other, due to the range $[0,1]$ being bounded.

Now we again want to define $\mathbb{P}(B|\mathcal{C})$ to be a function that inputs elements of X and outputs corresponding probability values in the interval $[0,1]$; but this time, it will only be defined up to $\mathbb{P}$-almost-sure equality, i.e. $\mathbb{P}(B|\mathcal{C})$ will be an element of the space $L(\mathbb{P}, [0,1])$.

Theorem. *Given an infinite collection $\mathcal{C}$ of test sets (countably infinite or uncountable), there exists an element of $L(\mathbb{P}, [0,1])$ that we will denote $\mathbb{P}(B|\mathcal{C})$, such that the following two statements hold:*

(1) *one can find finite collections $\mathcal{D} \subset \mathcal{C}$ for which $\mathbb{P}(B|\mathcal{D})$ is arbitrarily close to $\mathbb{P}(B|\mathcal{C})$;*

  → *more precisely: under the topology with which we have equipped $L(\mathbb{P}, [0,1])$, $\mathbb{P}(B|\mathcal{C})$ belongs to the closure of the set of those elements of $L(\mathbb{P}, [0,1])$ that are represented by functions of the form $\mathbb{P}(B|\mathcal{D})$ for a finite subcollection $\mathcal{D}$ of the collection $\mathcal{C}$;*

(2) *for any two finite collections $\mathcal{D}_1, \mathcal{D}_2 \subset \mathcal{C}$ with $\mathcal{D}_1 \subset \mathcal{D}_2$, we have*

$$\mathbb{E}\left[\Big(\mathbb{P}(B|\mathcal{D}_2) - \mathbb{P}(B|\mathcal{C})\Big)^2\right] \;\leq\; \mathbb{E}\left[\Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{C})\Big)^2\right].$$

Heuristically: We imagine testing a finite number of our infinitely many available test sets. Point (2) says that if we then test some further test sets from the available collection, our resulting posterior probability function will get closer – in terms of mean square error – to the element of $L(\mathbb{P}, [0,1])$ that we denote $\mathbb{P}(B|\mathcal{C})$. And Point (1) says that it is indeed possible to get arbitrarily close to this element by testing finitely many available test sets.

It is intuitive – and not hard to prove – that there is only one element $\mathbb{P}(B|\mathcal{C})$ of $L(\mathbb{P}, [0,1])$ fulfilling the description in the above theorem.

By applying Trivial Proposition 1 to finite subcollections of $\mathcal{C}$, one can obtain the following:

"Trivial Proposition" 2. *If $B \in \mathcal{C}$ then*

$$\mathbb{P}(B|\mathcal{C})(x) \;=\; \mathbb{1}_B(x) \;\; \mathbb{P}\text{-}a.s.(x).$$

(Here, "$\mathbb{P}$-a.s." stands for "$\mathbb{P}$-almost surely", meaning that, for any given representative function from X to $[0,1]$ representing the element $\mathbb{P}(B|\mathcal{C})$ of $L(\mathbb{P},[0,1])$, the set of points at which this function disagrees with $\mathbb{1}_B$ is infinitely unlikely to be landed in by the randomly selected point.)

Relationship between our definition and the standard definition. The standard definition of $\mathbb{P}(B|\mathcal{C})$ works with $\mathcal{C}$ being a "σ-algebra"; our definition of $\mathbb{P}(B|\mathcal{C})$, where $\mathcal{C}$ need not be a σ-algebra, coincides with the standard definition $\mathbb{P}(B|\sigma(\mathcal{C}))$, where $\sigma(\mathcal{C})$ denotes "the σ-algebra generated by $\mathcal{C}$".

Appendix of Part 1: Proofs.

These proofs will assume knowledge of basic measure theory. Here, "measurable" means "Borel-measurable" except where stated otherwise. The "probability distribution" $\mathbb{P}$ is a Borel probability measure on X.

Proof of Trivial Proposition 1

Assume $B \in \mathcal{C}$; then the definition of $\mathrm{Part}_X(\mathcal{C})$ yields that for any $A \in \mathrm{Part}_X(\mathcal{C})$, either $A \subset B$ or $A \subset X \setminus B$. Now given any $x \in X_{\mathbb{P},\mathcal{C}}$, let $A \in \mathrm{Part}_X(\mathcal{C})$ be such that $x \in A$. Then

$$\mathbb{P}(B|\mathcal{C})(x) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \begin{cases} 1 = \mathbb{1}_B(x) & \text{if } A \subset B \\ 0 = \mathbb{1}_B(x) & \text{if } A \subset X \setminus B. \end{cases}$$

Proof of Theorem

We will prove the Theorem with the inequality in property (2) strengthened to the following "Pythagorean" equality:

$$\mathbb{E}\left[\Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{D}_2)\Big)^2\right] + \mathbb{E}\left[\Big(\mathbb{P}(B|\mathcal{D}_2) - \mathbb{P}(B|\mathcal{C})\Big)^2\right] \;=\; \mathbb{E}\left[\Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{C})\Big)^2\right].$$

For a finite collection $\mathcal{D}$ of Borel subsets of X, write $\alpha(\mathcal{D})$ for the set of all unions of members of $\mathrm{Part}_X(\mathcal{D})$. Note that $\alpha(\mathcal{D})$ is an algebra of sets.

Lemma 1. *For any $\mathcal{D}_1 \subset \mathcal{D}$, $\mathrm{Part}_X(\mathcal{D}_1) \subset \alpha(\mathcal{D})$.*

Proof. Since $\mathcal{D}_1 \subset \mathcal{D} \subset \alpha(\mathcal{D})$ and $\alpha(\mathcal{D})$ is an algebra of sets, it follows immediately from the definition of $\mathrm{Part}_X(\mathcal{D}_1)$ that $\mathrm{Part}_X(\mathcal{D}_1) \subset \alpha(\mathcal{D})$. □

Now for a finite collection $\mathcal{D}$ of Borel subsets of X, for any bounded measurable $f\colon \tilde{X} \to \mathbb{R}$ where $\mathbb{P}(\tilde{X}) = 1$, define $\mathbb{E}[f|\mathcal{D}]\colon X_{\mathbb{P},\mathcal{D}} \to \mathbb{R}$ such that for each $x \in X_{\mathbb{P},\mathcal{D}}$,

$$x \in A \in \mathrm{Part}_X(\mathcal{D}) \implies \mathbb{E}[f|\mathcal{D}](x) = \frac{1}{\mathbb{P}(A)} \int_A f \, d\mathbb{P}.$$

Note that $\mathbb{P}(B|\mathcal{D}) = \mathbb{E}[\mathbb{1}_B|\mathcal{D}]$, since

$$\frac{1}{\mathbb{P}(A)} \int_A \mathbb{1}_B \, d\mathbb{P} \;=\; \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

for any A with $\mathbb{P}(A) > 0$.

Lemma 2. *We have the following properties:*

(A) Given $[a,b] \subset \mathbb{R}$, if the range of f is contained in $[a,b]$ then so is the range of $\mathbb{E}[f|\mathcal{D}]$.

(B) For any $A \in \alpha(\mathcal{D})$,

$$\int_A \mathbb{E}[f|\mathcal{D}]\, d\mathbb{P} = \int_A f\, d\mathbb{P}.$$

(C) In particular, taking $A = X$ in (B)*, we have $\mathbb{E}[\mathbb{E}[f|\mathcal{D}]] = \mathbb{E}[f]$.*

(D) For bounded measurable $f_1, f_2 \colon X \to \mathbb{R}$,

$$\mathbb{E}[f_1 + f_2|\mathcal{D}] = \mathbb{E}[f_1|\mathcal{D}] + \mathbb{E}[f_2|\mathcal{D}].$$

(E) For a bounded measurable function $c \colon X_1 \to \mathbb{R}$ where $X_{\mathbb{P},\mathcal{D}} \subset X_1$, if c is constant on each $A \in \mathrm{Part}_X(\mathcal{D})$ with $\mathbb{P}(A) > 0$ then

$$\mathbb{E}[c.f|\mathcal{D}](x) = c(x)\mathbb{E}[f|\mathcal{D}](x) \quad \forall\, x \in X_{\mathbb{P},\mathcal{D}}.$$

(F) For any $\mathcal{D}_1 \subset \mathcal{D}$,

$$\mathbb{E}[\,\mathbb{E}[f|\mathcal{D}]\,|\,\mathcal{D}_1] = \mathbb{E}[f|\mathcal{D}_1].$$

Proof. (A) Since $a \le f \le b$, we have $a\mathbb{P}(A) \le \int_A f\, d\mathbb{P} \le b\mathbb{P}(A)$ (for any measurable $A \subset X$), so the result is immediate. (B) Since $\mathrm{Part}_X(\mathcal{D})$ has finitely many members and they are disjoint, in order to show the desired equality for all $A \in \alpha(\mathcal{D})$, it is clearly sufficient just to show it for $A \in \mathrm{Part}_X(\mathcal{D})$ with $\mathbb{P}(A) > 0$; in this case,

$$\int_A \mathbb{E}[f|\mathcal{D}]\, d\mathbb{P} = \int_A \left(\frac{1}{\mathbb{P}(A)} \int_A f\, d\mathbb{P} \right) d\mathbb{P} = \mathbb{P}(A)\,.\,\frac{1}{\mathbb{P}(A)} \int_A f\, d\mathbb{P} = \int_A f\, d\mathbb{P}.$$

(C) and (D) are immediate. (E) For each $x \in X_{\mathbb{P},\mathcal{D}}$,

$$x \in A \in \mathrm{Part}_X(\mathcal{D}) \implies \mathbb{E}[c.f|\mathcal{D}](x) = \frac{1}{\mathbb{P}(A)} \int_A c.f\, d\mathbb{P} = \frac{1}{\mathbb{P}(A)} \int_A c(x) f\, d\mathbb{P} = c(x)\mathbb{E}[f|\mathcal{D}](x).$$

(F) For any $A \in \mathrm{Part}_X(\mathcal{D}_1)$ with $\mathbb{P}(A) > 0$, for any $x \in A$,

$$\begin{aligned} \mathbb{E}[\,\mathbb{E}[f|\mathcal{D}]\,|\,\mathcal{D}_1](x) &= \frac{1}{\mathbb{P}(A)} \int_A \mathbb{E}[f|\mathcal{D}]\, d\mathbb{P} \\ &= \frac{1}{\mathbb{P}(A)} \int_A f\, d\mathbb{P} \quad \text{by (B) and Lemma 1} \\ &= \mathbb{E}[f|\mathcal{D}_1](x). \end{aligned}$$

□

Since X is a separable metric space, the Borel σ-algebra is countably generated, and so it is known that for $p \in [1,\infty)$ the space $L^p(\mathbb{P})$ of measurable functions $f \colon X \to \mathbb{R}$ with $\mathbb{E}[|f|^p] < \infty$, identified up to $\mathbb{P}$-almost sure equality, is a separable complete metric space under the distance function $d_p(f_1,f_2) := \mathbb{E}[|f_1 - f_2|^p]^{1/p}$. In our notation, we will not generally distinguish between an exactly-defined function f and the element of $L^p(\mathbb{P})$ that it represents. Now $L(\mathbb{P},[0,1])$ is a closed subset of $L^p(\mathbb{P})$, and so $L(\mathbb{P},[0,1])$ is a separable complete metric space under d_p. For any measurable $A_1, A_2 \subset X$, we define

$$A_1 \triangle A_2 = \{x \in X : \mathbb{1}_{A_1}(x) \neq \mathbb{1}_{A_2}(x)\} = (A_1 \setminus A_2) \cup (A_2 \setminus A_1).$$

Note that $d_p(\mathbb{1}_{A_1}, \mathbb{1}_{A_2}) = \mathbb{P}(A_1 \triangle A_2)^{\frac{1}{p}}$. For $p = 2$, we equip $L^2(\mathbb{P})$ with an inner product, namely $\langle f, g \rangle_{L^2} = \mathbb{E}[fg]$; the metric d_2 is precisely the metric induced by this inner product.

Lemma 3. *Given three finite collections $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \mathcal{D}_3$ and a measurable $B \subset X$, we have that $\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{D}_2)$ and $\mathbb{P}(B|\mathcal{D}_2) - \mathbb{P}(B|\mathcal{D}_3)$ are L^2-orthogonal to each other.*

Proof. Two functions f and g with $\mathbb{E}[fg|\mathcal{D}_2] = 0$ will have $\mathbb{E}[fg] = 0$ by Lemma 2(C), i.e. they will be orthogonal to each other. Now

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{D}_2)\Big)\Big(\mathbb{P}(B|\mathcal{D}_2) - \mathbb{P}(B|\mathcal{D}_3)\Big)\Big|\mathcal{D}_2\Big] \\
&= \Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{D}_2)\Big)\mathbb{E}\Big[\mathbb{P}(B|\mathcal{D}_2) - \mathbb{P}(B|\mathcal{D}_3)\Big|\mathcal{D}_2\Big] \quad \text{by Lemma 2(E)} \\
&= \Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{D}_2)\Big)\Big(\mathbb{P}(B|\mathcal{D}_2) - \mathbb{E}[\mathbb{P}(B|\mathcal{D}_3)|\mathcal{D}_2]\Big) \quad \text{by Lemma 2(D,E)} \\
&= \Big(\mathbb{P}(B|\mathcal{D}_1) - \mathbb{P}(B|\mathcal{D}_2)\Big)\Big(\mathbb{P}(B|\mathcal{D}_2) - \mathbb{P}(B|\mathcal{D}_2)\Big) \quad \text{by Lemma 2(F)} \\
&= 0.
\end{aligned}$$

□

Lemma 4. *Given a finite collection $\mathcal{D}$ and a measurable $C \subset X$, for any measurable $B \subset X$ we have*

$$d_2\Big(\mathbb{P}(B|\mathcal{D})\,,\,\mathbb{P}(B|\mathcal{D}\cup\{C\})\Big) \le \tfrac{\sqrt{5}}{2}\min_{E\in\alpha(\mathcal{D})}\mathbb{P}(C\triangle E)^{\frac{1}{4}}.$$

To prove this, we start with the following simple fact.

Lemma 5. *In the setting of Lemma 4, let us write $\eta := \min_{E\in\alpha(\mathcal{D})}\mathbb{P}(C\triangle E)$, and for each $A \in \mathrm{Part}_X(\mathcal{D})$ define $\phi(A) \subset A$ by*

$$\phi(A) := \begin{cases} A\cap C & \text{if } \mathbb{P}(A\cap C) \le \frac{1}{2}\mathbb{P}(A) \\ A\setminus C & \text{if } \mathbb{P}(A\cap C) > \frac{1}{2}\mathbb{P}(A). \end{cases}$$

Then

$$\eta = \mathbb{P}\left(\bigcup_{A\in\mathrm{Part}_X(\mathcal{D})}\phi(A)\right) = \sum_{A\in\mathrm{Part}_X(\mathcal{D})}\mathbb{P}(\phi(A)).$$

Proof. The elements E of $\alpha(\mathcal{D})$ can be identified precisely by specifying whether each member A of the partition $\mathrm{Part}_X(\mathcal{D})$ is a subset of E or a subset of $X\setminus E$, and this is in turn equivalent to specifying whether the part of $C\triangle E$ contained in A is $A\setminus C$ or $A\cap C$; and $\phi(A)$ is defined to be either $A\setminus C$ or $A\cap C$ in such a way that if $A\cap C$ and $A\setminus C$ have different measure then $\phi(A)$ is the one with the smaller measure. □

Proof of Lemma 4. We continue with the notations introduced in Lemma 5. Let

$$\mathcal{A} := \{A \in \mathrm{Part}_X(\mathcal{D}) : \mathbb{P}(\phi(A)) > \sqrt{\eta}\,\mathbb{P}(A)\},$$

and let $\mathcal{A}^c := \mathrm{Part}_X(\mathcal{D})\setminus\mathcal{A}$. By Lemma 5,

$$\sum_{A\in\mathcal{A}}\mathbb{P}(\phi(A)) \le \eta.$$

If $\eta = 0$ then it follows that $\mathbb{P}(\phi(A)) = 0$ for all $A \in \mathcal{A}$ and hence (by virtue of the definition of $\mathcal{A}$) that $\mathcal{A} = \emptyset$; and if $\eta > 0$ then, again using the definition of $\mathcal{A}$, it follows that $\sum_{A\in\mathcal{A}}\sqrt{\eta}\,\mathbb{P}(A) \le \eta$, and we can divide both sides by $\sqrt{\eta}$; so in either case, it follows that

$$\sum_{A\in\mathcal{A}}\mathbb{P}(A) \le \sqrt{\eta}.$$

Now for each $A \in \mathrm{Part}_X(\mathcal{D})$:

- if $A \subset C$ or $A \subset X \setminus C$ then A is a member of $\mathrm{Part}_X(\mathcal{D} \cup \{C\})$;
- if A intersects both C and $X \setminus C$ then A is the union of two members of $\mathrm{Part}_X(\mathcal{D} \cup \{C\})$, namely $A \cap C$ and $A \setminus C$; and furthermore, if $0 < \mathbb{P}(A \cap C) < 1$ then

$$\mathbb{P}(B|A) = \mathbb{P}(C|A)\mathbb{P}(B|A \cap C) + \big(1 - \mathbb{P}(C|A)\big)\mathbb{P}(B|A \setminus C). \tag{1}$$

Consequently, writing

$$\begin{aligned} v(A) &:= \int_A \left(\mathbb{P}(B|\mathcal{D} \cup \{C\}) - \mathbb{P}(B|\mathcal{D})\right)^2 \, d\mathbb{P} \\ &= \int_{A \cap C} \left(\mathbb{P}(B|\mathcal{D} \cup \{C\}) - \mathbb{P}(B|\mathcal{D})\right)^2 \, d\mathbb{P} + \int_{A \setminus C} \left(\mathbb{P}(B|\mathcal{D} \cup \{C\}) - \mathbb{P}(B|\mathcal{D})\right)^2 \, d\mathbb{P}, \end{aligned}$$

we have that

- if $0 < \mathbb{P}(A \cap C) < 1$ then

$$\begin{aligned} v(A) &= \mathbb{P}(A \cap C)\big(\mathbb{P}(B|A \cap C) - \mathbb{P}(B|A)\big)^2 + \mathbb{P}(A \setminus C)\big(\mathbb{P}(B|A \setminus C) - \mathbb{P}(B|A)\big)^2 \\ &= \mathbb{P}(A)\Big[\mathbb{P}(C|A)\big(\mathbb{P}(B|A \cap C) - \mathbb{P}(B|A)\big)^2 + \big(1 - \mathbb{P}(C|A)\big)\big(\mathbb{P}(B|A \setminus C) - \mathbb{P}(B|A)\big)^2\Big] \\ &= \mathbb{P}(A)\Big(\mathbb{P}(B|A \cap C) - \mathbb{P}(B|A \setminus C)\Big)^2 \mathbb{P}(C|A)\big(1 - \mathbb{P}(C|A)\big) \quad \text{using (1);} \end{aligned}$$

- if $\mathbb{P}(A \cap C)$ is 0 or 1, then $v(A) = 0$.

In the former case, we have that $v(A) \leq \frac{1}{4}\mathbb{P}(A)$ (since the maximum value of $\lambda \mapsto \lambda(1-\lambda)$ on $[0,1]$ is $\frac{1}{4}$), and that if $A \in \mathcal{A}^c$ then

$$v(A) \leq \mathbb{P}(A) \min(\, \mathbb{P}(C|A)\,,\, 1 - \mathbb{P}(C|A)\,) = \mathbb{P}(\phi(A)) \leq \sqrt{\eta}\,\mathbb{P}(A).$$

Hence

$$\begin{aligned} d_2\Big(\mathbb{P}(B|\mathcal{D})\,,\, \mathbb{P}(B|\mathcal{D} \cup \{C\})\Big)^2 &= \sum_{A \in \mathrm{Part}_X(\mathcal{D})} v(A) \\ &\leq \left(\sum_{A \in \mathcal{A}} \tfrac{1}{4}\mathbb{P}(A)\right) + \left(\sum_{A \in \mathcal{A}^c} \sqrt{\eta}\,\mathbb{P}(A)\right) \\ &\leq \tfrac{1}{4}\sqrt{\eta} + \sqrt{\eta} \\ &= \tfrac{5}{4}\sqrt{\eta}. \end{aligned}$$

□

As an immediate corollary of Lemma 4, we have the following.

Corollary 6. *For all $\varepsilon > 0$ and $n \in \mathbb{N}$ there exists $\delta(\varepsilon, n) > 0$ such that for any measurable $B \subset X$ and finite collections $\mathcal{D}, \mathcal{D}'$ with $|\mathcal{D}'| = n$, if*

$$\max_{C \in \mathcal{D}'} \min_{E \in \alpha(\mathcal{D})} \mathbb{P}(C \triangle E) \leq \delta$$

then

$$d_2\Big(\mathbb{P}(B|\mathcal{D})\,,\, \mathbb{P}(B|\mathcal{D} \cup \mathcal{D}')\Big) \leq \varepsilon.$$

We now prove the Theorem. Since $L(\mathbb{P},[0,1])$ is separable, we can find a countable subset $\{C_n\}_{n\geq 1}$ of $\mathcal{C}$ that is dense in $\mathcal{C}$ (i.e. for every $C\in\mathcal{C}$, $\inf_{n\geq 1}\mathbb{P}(C_n\triangle C)=0$). Let $\tilde{\mathcal{D}}_n=\{C_1,\ldots,C_n\}$ for each n. Then by Lemma 3, for all $n\geq m\geq 1$ we have

$$d_2\Big(\mathbb{P}(B|\tilde{\mathcal{D}}_1)\,,\,\mathbb{P}(B|\tilde{\mathcal{D}}_n)\Big)^2=d_2\Big(\mathbb{P}(B|\tilde{\mathcal{D}}_1)\,,\,\mathbb{P}(B|\tilde{\mathcal{D}}_m)\Big)^2+d_2\Big(\mathbb{P}(B|\tilde{\mathcal{D}}_m)\,,\,\mathbb{P}(B|\tilde{\mathcal{D}}_n)\Big)^2.$$

Since $L(\mathbb{P},[0,1])$ is d_2-bounded, it follows that the sequence $\big(\mathbb{P}(B|\tilde{\mathcal{D}}_n)\big)_{n\geq 1}$ is d_2-Cauchy. Since $L(\mathbb{P},[0,1])$ is d_2-complete, this sequence converges to a limit that we denote $\mathbb{P}(B|\mathcal{C})$.[1] Obviously $\mathbb{P}(B|\mathcal{C})$ fulfils property (1) of the Theorem. Now take any finite $\mathcal{D}_1\subset\mathcal{D}_2\subset\mathcal{C}$, and for all $n\geq 3$ define $\mathcal{D}_n=\mathcal{D}_2\cup\tilde{\mathcal{D}}_n$. Since $\mathcal{D}_2\subset\mathcal{C}$ and $\{C_n\}_{n\geq 1}$ is dense in $\mathcal{C}$, applying Corollary 6 gives that

$$d_2\Big(\mathbb{P}(B|\mathcal{D}_n)\,,\,\mathbb{P}(B|\tilde{\mathcal{D}}_n)\Big)\to 0\quad\text{as } n\to\infty,$$

and hence

$$\mathbb{P}(B|\mathcal{D}_n)\to\mathbb{P}(B|\mathcal{C})\quad\text{as } n\to\infty.$$

Lemma 3 gives that

$$d_2\Big(\mathbb{P}(B|\mathcal{D}_1)\,,\,\mathbb{P}(B|\mathcal{D}_n)\Big)^2=d_2\Big(\mathbb{P}(B|\mathcal{D}_1)\,,\,\mathbb{P}(B|\mathcal{D}_2)\Big)^2+d_2\Big(\mathbb{P}(B|\mathcal{D}_2)\,,\,\mathbb{P}(B|\mathcal{D}_n)\Big)^2$$

for each $n\geq 2$, and taking the limit as $n\to\infty$ gives the desired result.

Proof that $\mathbb{P}(B|\mathcal{C})$ is unique

Suppose we have distinct $f_1,f_2\in L(\mathbb{P},[0,1])$ both fulfilling the description of $\mathbb{P}(B|\mathcal{C})$ in the Theorem. Letting $\varepsilon:=\frac{1}{2}d_2(f_1,f_2)>0$, we can find $\mathcal{D},\mathcal{D}'\subset\mathcal{C}$ such that

$$d_2\big(\mathbb{P}(B|\mathcal{D})\,,\,f_1\big)<\varepsilon$$
$$d_2\big(\mathbb{P}(B|\mathcal{D}')\,,\,f_2\big)<\varepsilon.$$

It then follows that

$$d_2\big(\mathbb{P}(B|\mathcal{D}\cup\mathcal{D}')\,,\,f_1\big)<\varepsilon$$
$$d_2\big(\mathbb{P}(B|\mathcal{D}\cup\mathcal{D}')\,,\,f_2\big)<\varepsilon,$$

and hence, by the triangle inequality, $d_2(f_1,f_2)<2\varepsilon=d_2(f_1,f_2)$, giving a contradiction.

Proof of Trivial Proposition 2

Assume $B\in\mathcal{C}$. On the basis of property (1) in the Theorem, let $(\mathcal{D}_n)$ be a sequence of finite subcollections of $\mathcal{C}$ such that $\mathbb{P}(B|\mathcal{D}_n)$ converges to $\mathbb{P}(B|\mathcal{C})$ as $n\to\infty$. Due to property (2) in the Theorem, for each n, we have that

$$d_2\big(\mathbb{P}(B|\mathcal{D}_n\cup\{B\})\,,\,\mathbb{P}(B|\mathcal{C})\big)\leq d_2\big(\mathbb{P}(B|\mathcal{D}_n)\,,\,\mathbb{P}(B|\mathcal{C})\big),$$

and hence $\mathbb{P}(B|\mathcal{D}_n\cup\{B\})$ converges to $\mathbb{P}(B|\mathcal{C})$ as $n\to\infty$. But by Trivial Proposition 1, $\mathbb{P}(B|\mathcal{D}_n\cup\{B\})\overset{\mathbb{P}\text{-a.s.}}{=}\mathbb{1}_B$; i.e., in $L(\mathbb{P},[0,1])$ the sequence $\mathbb{P}(B|\mathcal{D}_n\cup\{B\})$ is simply the constant sequence $\mathbb{1}_B$.

[1] It fact, it turns out ("Lévy's Upward Theorem") that we have not merely convergence in the topology of $L(\mathbb{P},[0,1])$ but $\mathbb{P}$-almost-sure convergence.

Proof that $\mathbb{P}(B|\mathcal{C})$ coincides with the classical definition of $\mathbb{P}(B|\sigma(\mathcal{C}))$

The classical definition of $\mathbb{P}(B|\mathcal{C})$ assumes that $\mathcal{C}$ is a σ-algebra (while B may be any Borel set), and defines $\mathbb{P}(B|\mathcal{C})$ to be the unique element of $L(\mathbb{P},[0,1])$ that admits a $\mathcal{C}$-measurable version and has

$$\mathbb{P}(B\cap C)=\int_C \mathbb{P}(B|\mathcal{C})\,d\mathbb{P}$$

for all $C\in\mathcal{C}$.

From now on, we return to interpreting "$\mathbb{P}(B|\mathcal{C})$" according to our definition, without assuming that $\mathcal{C}$ is a σ-algebra. If $\mathcal{C}$ is a σ-algebra, then given any $f\in L^1(\mathbb{P})$ we will write $\overline{\mathbb{E}[f|\mathcal{C}]}\in L^1(\mathbb{P})$ for the classical interpretation of "$\mathbb{E}[f|\mathcal{C}]$", i.e. $\overline{\mathbb{E}[f|\mathcal{C}]}$ has a $\mathcal{C}$-measurable version and

$$\int_C f\,d\mathbb{P}=\int_C \overline{\mathbb{E}[f|\mathcal{C}]}\,d\mathbb{P}$$

for all $C\in\mathcal{C}$.

Lemma 7. $\mathbb{P}(B|\mathcal{C})=\mathbb{P}(B|\sigma(\mathcal{C}))$.

Proof. Let $(\tilde{\mathcal{D}}_n)_{n\geq 1}$ be as in the proof of the Theorem, and let $\mathcal{E}=\bigcup_{n=1}^{\infty}\alpha(\tilde{\mathcal{D}}_n)$. Note that $\mathcal{E}$ is an algebra of sets. Let $\bar{\mathcal{E}}$ be the "closure" of $\mathcal{E}$, i.e.

$$\bar{\mathcal{E}}=\left\{\text{measurable } C\subset X:\inf_{E\in\mathcal{E}}\mathbb{P}(C\triangle E)=0\right\}.$$

By definition of $(\tilde{\mathcal{D}}_n)_{n\geq 1}$, we have that $\mathcal{C}\subset\bar{\mathcal{E}}$. We next want to show that $\sigma(\mathcal{C})\subset\bar{\mathcal{E}}$, and so for this it will be sufficient to show that $\bar{\mathcal{E}}$ is a σ-algebra. Obviously $\emptyset,X\in\bar{\mathcal{E}}$. For any $C\in\bar{\mathcal{E}}$, if (E_n) is a sequence in $\mathcal{E}$ such that $\mathbb{P}(C\triangle E_n)\to 0$, then $(X\setminus E_n)$ is also a sequence in $\mathcal{E}$ and we have

$$(X\setminus C)\triangle(X\setminus E_n)=C\triangle E_n;$$

so $X\setminus C\in\bar{\mathcal{E}}$. For any $C_1,C_2\in\bar{\mathcal{E}}$ and any $\varepsilon>0$, if we take $E_1,E_2\in\mathcal{E}$ with $\mathbb{P}(C_i\triangle E_i)<\frac{\varepsilon}{2}$ for $i=1,2$, then since

$$(C_1\cup C_2)\triangle(E_1\cup E_2)\ \subset\ (C_1\triangle E_1)\cup(C_2\triangle E_2),$$

we have $\mathbb{P}((C_1\cup C_2)\triangle(E_1\cup E_2))<\varepsilon$, and $E_1\cup E_2\in\mathcal{E}$; so $C_1\cup C_2\in\bar{\mathcal{E}}$. So $\bar{\mathcal{E}}$ is an algebra. For any increasing sequence $(C_n)_{n\geq 1}$ in $\bar{\mathcal{E}}$, letting $C_\infty=\bigcup_{n=1}^{\infty}C_n$, we have that $\mathbb{P}(C_\infty\triangle C_n)=\mathbb{P}(C_\infty\setminus C_n)\to 0$ as $n\to\infty$; but $\bar{\mathcal{E}}$ is closed under limits of convergent sequences (in terms of $\mathbb{P}(\cdot\triangle\cdot)$ tending to 0), and hence $C_\infty\in\bar{\mathcal{E}}$. So the algebra $\bar{\mathcal{E}}$ is indeed a σ-algebra. Thus, in particular, $\sigma(\mathcal{C})\subset\bar{\mathcal{E}}$. So $\mathcal{E}$ is a countable subset of $\sigma(\mathcal{C})$ that is "dense" in $\sigma(\mathcal{C})$, and thus, as in the proof of the Theorem, we have

$$\mathbb{P}(B|\alpha(\tilde{\mathcal{D}}_n))\to\mathbb{P}(B|\sigma(\mathcal{C}))\quad\text{as } n\to\infty.$$

But we also have that for each n, $\mathrm{Part}_X(\alpha(\tilde{\mathcal{D}}_n))=\mathrm{Part}_X(\tilde{\mathcal{D}}_n)$, and so $\mathbb{P}(B|\alpha(\tilde{\mathcal{D}}_n))$ is equal to $\mathbb{P}(B|\tilde{\mathcal{D}}_n)$, which converges to $\mathbb{P}(B|\mathcal{C})$ as $n\to\infty$. □

Lemma 8. *For all $C\in\mathcal{C}$, $\mathbb{P}(B\cap C)=\int_C\mathbb{P}(B|\mathcal{C})\,d\mathbb{P}$.*

Proof. Fix $C\in\mathcal{C}$, and let $(\tilde{\mathcal{D}}_n)_{n\geq 1}$ be as in the proof of the Theorem. Let $(\tilde{C}_n)_{n\geq 1}$ be a sequence with $\tilde{C}_n\in\tilde{\mathcal{D}}_n$ for each n, such that $\mathbb{P}(\tilde{C}_n\triangle C)\to 0$ as $n\to\infty$. Lemma 2(B) gives that for each n,

$$\mathbb{P}(B\cap\tilde{C}_n)=\int_{\tilde{C}_n}\mathbb{P}(B|\tilde{\mathcal{D}}_n)\,d\mathbb{P}.$$

Since $\mathbb{P}(B|\tilde{\mathcal{D}}_n)\to\mathbb{P}(B|\mathcal{C})$ and $\mathbb{1}_{\tilde{C}_n}\to\mathbb{1}_C$ as $n\to\infty$, it is not hard to show that we can obtain the desired result by taking the limit as $n\to\infty$ in the above equality. □

In view of Lemmas 7 and 8, it remains only to justify that $\mathbb{P}(B|\mathcal{C})$ has a $\sigma(\mathcal{C})$-measurable version. Since convergence in probability implies almost-sure convergence of a subsequence, we can find a sequence $(\mathcal{D}_n)$ of finite subcollections of $\mathcal{C}$ such that $\mathbb{P}(B|\mathcal{D}_n)$ converges $\mathbb{P}$-almost surely to $\mathbb{P}(B|\mathcal{C})$ as $n \to \infty$. Since $\mathbb{P}(B|\mathcal{D}_n)$ is constant on each member of $\mathrm{Part}_X(\mathcal{D}_n)$, we have in particular that $\mathbb{P}(B|\mathcal{D}_n)$ is $\sigma(\mathcal{C})$-measurable. Hence there is a Borel-measurable set $X_0 \subset X$ with $\mathbb{P}(X_0) = 1$ such that $\mathbb{P}(B|\mathcal{C})\big|_{X_0}$ is measurable with respect to the induced σ-algebra from $\sigma(\mathcal{C})$ on X_0. Now let g be a $\sigma(\mathcal{C})$-measurable version of $\overline{\mathbb{E}\big[\,\mathbb{P}(B|\mathcal{C})\,\big|\sigma(\mathcal{C})\big]}$. Then for all $C \in \sigma(\mathcal{C})$,

$$\int_{C\cap X_0} g\,d\mathbb{P} = \int_C g\,d\mathbb{P} = \int_C \mathbb{P}(B|\mathcal{C})\,d\mathbb{P} = \int_{C\cap X_0} \mathbb{P}(B|\mathcal{C})\,d\mathbb{P},$$

and so $\mathbb{P}(B|\mathcal{C})$ agrees with g $\mathbb{P}$-almost everywhere in X_0, and hence in X.

Part 2 (Thursday 9th February 2023)

(We continue with our Standing Assumption that all subsets of X that we mention are Borel sets.)

So far, we have imagined that there is a particular set $B \subset X$ of interest, and asked about the posterior probability that the randomly selected point x lies in B given observability of $\mathcal{C}$.

Now let us suppose that there is not a pre-identified region of the state space X that is of special interest: rather, we simply want to know

> **"what is the posterior probability distribution of the randomly selected point x, given observability of $\mathcal{C}$?"**

such that someone could then ask me about *any* region B of the state space and, *on the basis of my posterior probability distribution for* x, I would accordingly be able to tell them the posterior probability that x is in B. (Note that, by definition, the "*prior probability distribution* of x" is simply $\mathbb{P}$.)

By way of notation: for any $A \subset X$ with $\mathbb{P}(A) > 0$, the "probability distribution of x conditional on the observation that $x \in A$" is denoted $\mathbb{P}_A$; to be precise, this is defined such that for each $B \subset X$,

$$\mathbb{P}_A(B) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

Now once again, we want to represent the answer to the above question as a function $\mathbb{P}_{\mathcal{C}}$, whose inputs are elements of X and whose outputs are probability distributions defined over X, where if we input the invisible true value of x, the corresponding output is the posterior probability distribution of x given observability of $\mathcal{C}$.

Definition. If $\mathcal{C}$ is finite, define

$$\mathbb{P}_{\mathcal{C}} \colon X_{\mathbb{P},\mathcal{C}} \to \{\text{probability distributions on } X\}$$

such that for every $A \in \mathrm{Part}_X(\mathcal{C})$ with $\mathbb{P}(A) > 0$,

$$\forall\, x \in A,\ \mathbb{P}_{\mathcal{C}} = \mathbb{P}_A.$$

In other words: $\forall\, x \in X_{\mathbb{P},\mathcal{C}}$, for any $B \subset X$,

$$\mathbb{P}_{\mathcal{C}}(x)\big(B\big) = \mathbb{P}(B|\mathcal{C})(x). \tag{2}$$

We now consider the case that $\mathcal{C}$ is infinite. Recall that in this case, we defined in Part 1 an element $\mathbb{P}(B|\mathcal{C})$ of the space $L(\mathbb{P}, [0,1])$ of functions $X \to [0,1]$ identified up to $\mathbb{P}$-almost sure equality.

Theorem.[2] *Given an infinite collection $\mathcal{C}$, there exists—unique up to $\mathbb{P}$-almost sure equality—a function*

$$\mathbb{P}_{\mathcal{C}} \colon X \to \{\text{probability distributions on } X\}$$

with the property that for each $B \subset X$, the map $x \mapsto \mathbb{P}_{\mathcal{C}}(x)\big(B\big)$ is a version of $\mathbb{P}(B|\mathcal{C})$

$\to$ *i.e. with the property that for each $B \subset X$, Eq. (2) $\mathbb{P}$-almost surely holds.*

The "unique up to $\mathbb{P}$-almost sure equality" means that if we have two functions P_1 and P_2 fulfilling the description in this theorem, then it is infinitely unlikely that the randomly selected point x will be such that $P_1(x)$ and $P_2(x)$ are distinct probability distributions.

Ergodic Decomposition Theorem

Let $T \colon X \to X$ be a Borel-measurable map. (Again, this means that T is "not too horrible"; any self-map of X that you might ever want to consider in practice will fulfil this.)

We will define "T-invariant probability distributions" and "T-invariant subsets of X".

Definition. We say that a probability distribution $\mathbb{P}$ on X is *T-invariant* if for an X-valued random variable R with probability distribution $\mathbb{P}$, the random variable $T(R)$ also has probability distribution $\mathbb{P}$.

Definition. We say that a set $A \subset X$ is *T-invariant* if A is a Borel set and for all $x \in X$,

$$x \in A \;\Leftrightarrow\; T(x) \in A.$$

(Despite our Standing Assumption, it is worth very explicitly including the "Borel set" property as part of the definition, as failing to have this would considerably affect the concept of "ergodicity".)

Note that

- if A is T-invariant then so is $X \setminus A$;
- if A_1, A_2 are T-invariant then so are $A_1 \cup A_2$ and $A_1 \cap A_2$.

T-invariant probability distributions and T-invariant sets are related by the following result that is not hard to prove:

Proposition. *If $\mathbb{P}$ is T-invariant and $A \subset X$ is a T-invariant set with $\mathbb{P}(A) > 0$, then $\mathbb{P}_A$ is T-invariant.*

Now if $\mathbb{P}$ is T-invariant and $\mathcal{C}$ is a finite collection of T-invariant sets, then the "law of total probability" gives

$$\mathbb{P}(\cdot) = \sum_{\substack{A \in \mathrm{Part}_X(\mathcal{C}),\\ \mathbb{P}(A)>0}} \mathbb{P}(A)\mathbb{P}_A(\cdot), \tag{3}$$

[2] This is a well-known result (Theorem 33.3 of P. Billingsley, *Probability and Measure*, 3rd Edition, 1995); it is a case of the *disintegration theorem*.

where in the summand, $\mathbb{P}(A)$ is a scalar coefficient, while by the above Proposition, $\mathbb{P}_A(\cdot)$ is a T-invariant probability distribution. Thus, (3) serves as a "*decomposition of the T-invariant probability distribution $\mathbb{P}$ into T-invariant components*".

Now the "sum" in (3) could have only one term in it. Specifically, this happens if and only if every $C \in \mathcal{C}$ has either $\mathbb{P}(C) = 0$ or $\mathbb{P}(C) = 1$. In this case, we say that the decomposition (3) is a "*trivial decomposition*".

Definition. We say that a T-invariant probability distribution $\mathbb{P}$ is *T-ergodic* if "it does not admit a non-trivial decomposition of the form (3)", i.e. if every T-invariant set $C \subset X$ has $\mathbb{P}(C) = 0$ or $\mathbb{P}(C) = 1$.

Now let us point out that in general, for a T-invariant probability distribution $\mathbb{P}$ and a finite collection $\mathcal{C}$ of T-invariant sets, the "invariant components" $\mathbb{P}_A$ in the decomposition (3) need not be T-ergodic, i.e. *they may themselves be "further decomposable" using a "finer" collection $\mathcal{C}' \supset \mathcal{C}$ of T-invariant sets.* This naturally leads to the following question: what about if, instead of taking a finite collection of T-invariant sets, we choose $\mathcal{C}$ to be the *entire* collection of all T-invariant sets.

Theorem (Ergodic Decomposition Theorem)**.** *Let $\mathcal{C}$ be the set of all T-invariant sets. If $\mathbb{P}$ is T-invariant, then for $\mathbb{P}$-almost every $x \in X$, $\mathbb{P}_{\mathcal{C}}(x)$ is T-ergodic.*

The "$\mathbb{P}$-almost every" is included simply because $\mathbb{P}_{\mathcal{C}}$ was only defined uniquely up to $\mathbb{P}$-almost-sure equality; it would be possible to choose a version of $\mathbb{P}_{\mathcal{C}}$ such that $\mathbb{P}_{\mathcal{C}}(x)$ is T-ergodic for *every* $x \in X$.

Now the proof of the Ergodic Decomposition Theorem has two parts:

① [the "relatively easy" part] Show that $\mathbb{P}_{\mathcal{C}}(x)$ is T-invariant for $\mathbb{P}$-almost all x.

② [the harder part] Given ①, show that $\mathbb{P}_{\mathcal{C}}(x)$ is T-ergodic for $\mathbb{P}$-almost all x.

Part ② is, as we have just indicated, the harder part – *except* that in some references it appears to be very easy:

> **"**by Trivial Proposition 2, for $\mathbb{P}$-almost every x we have
>
> $$\forall\, B \in \mathcal{C},\ \mathbb{P}_{\mathcal{C}}(x)\big(B\big) = \mathbb{1}_B(x)$$
>
> and so $\mathbb{P}_{\mathcal{C}}(x)$ assigns 0 or 1 to every T-invariant set B.**"**

But there is an error in this "proof"; contained within the quotation marks is a wrong statement for which it is possible to find counterexamples.

Part 3 (Thursday 16th February 2023)

In fact, not only is it *possible* to find counterexamples, but the situation in which what is contained within the quotation marks is correct is quite degenerate:

Exercise. Suppose T is invertible. Show that if the statements contained within the quotation marks are correct, then $\mathbb{P}$-almost every $x \in X$ is a fixed or periodic point of T.

(The condition that T is invertible is not actually needed; and also, regardless of whether T is invertible, the "if" can be strengthened to "if and only if". But all this is somewhat harder to show, while the Exercise as written above is not very difficult.)

The flaw in the "proof" is that it is only for *each individual* $B \in \mathcal{C}$ that we can say that for $\mathbb{P}$-almost every x, $\mathbb{P}_{\mathcal{C}}(x)\big(B\big) = \mathbb{1}_B(x)$; the collection $\mathcal{C}$ can be infinite – indeed, it can easily be a "continuum" of sets. Given a continuum for each member of which some statement holds $\mathbb{P}$-almost surely, we cannot necessarily conclude that $\mathbb{P}$-almost surely, all members of the continuum satisfy the statement.

Let us now formulate the issue is general terms outside the specific setting of the Ergodic Decomposition Theorem.

Generalising beyond the dynamics setting

Returning to the general setting of a probability distribution $\mathbb{P}$ on X and a collection $\mathcal{C}$ of subsets of X:

Definition. We say that $\mathcal{C}$ *is regular with respect to* $\mathbb{P}$ if $\mathbb{P}$-almost every $x \in X$ has that for all $B \in \mathcal{C}$, $\mathbb{P}_{\mathcal{C}}(x)\big(B\big) = \mathbb{1}_B(x)$.

By Trivial Proposition 2, if $\mathcal{C}$ is finite (or countably infinite) then it is regular with respect to $\mathbb{P}$.

Definition. We say that $\mathcal{C}$ *is self-conditionally trivial with respect to* $\mathbb{P}$ if $\mathbb{P}$-almost every $x \in X$ has that for all $B \in \mathcal{C}$, $\mathbb{P}_{\mathcal{C}}(x)\big(B\big) \in \{0, 1\}$.

So, once we are given Part ① of the proof of the Ergodic Decomposition Theorem, the Ergodic Decomposition Theorem then says precisely that the set of all T-invariant sets is self-conditionally trivial with respect to the T-invariant probability distribution $\mathbb{P}$.

Obviously, if $\mathcal{C}$ is regular then it is self-conditionally trivial, but the converse does not hold. The above "proof" of the Ergodic Decomposition Theorem falsely assumes that any collection $\mathcal{C}$ is automatically regular. Let us now give an example of a situation in which regularity fails: Take $X = [0, 1]$, with $\mathbb{P}$ the uniform distribution, and $\mathcal{C} = \{\{x\} : x \in [0, 1]\}$. For any finite $\mathcal{D} \subset \mathcal{C}$, writing $\mathcal{D} = \{\{x_1\}, \ldots, \{x_n\}\}$, we have that $\mathrm{Part}_X(\mathcal{D}) = \mathcal{D} \cup \{[0, 1] \setminus \{x_1, \ldots, x_n\}\}$, and hence in particular, $X_{\mathbb{P},\mathcal{C}} = [0, 1] \setminus \{x_1, \ldots, x_n\}$. So for each $B \subset X$, we have

$$\mathbb{P}(B|\mathcal{D})\colon [0, 1] \setminus \{x_1, \ldots, x_n\} \to [0, 1]$$
$$\mathbb{P}(B|\mathcal{D})(x) = \frac{\mathbb{P}(B \setminus \{x_1, \ldots, x_n\})}{\mathbb{P}([0, 1] \setminus \{x_1, \ldots, x_n\})} = \frac{\mathbb{P}(B)}{1} = \mathbb{P}(B).$$

It follows that $\mathbb{P}(B|\mathcal{C})(x) \overset{\mathbb{P}\text{-a.s.}(x)}{=} \mathbb{P}(B)$. As this holds for every B, we see from the definition of $\mathbb{P}_{\mathcal{C}}$ that $\mathbb{P}_{\mathcal{C}}(x) \overset{\mathbb{P}\text{-a.s.}(x)}{=} \mathbb{P}$. Now for $\mathbb{P}$-almost every $x \in X$, we have that $\{x\} \in \mathcal{C}$ and yet

$$\mathbb{P}_{\mathcal{C}}(x)\big(\{x\}\big) = \mathbb{P}\big(\{x\}\big) = 0 \neq \mathbb{1}_{\{x\}}(x).$$

So we have given an simple example showing that one cannot assume that any collection $\mathcal{C}$ is regular. But nonetheless, as we have said, the Ergodic Decomposition Theorem still says that the collection $\mathcal{C}$ consisting of all T-invariant sets is self-conditionally trivial.

This is curious, as – at least at first sight – self-conditional triviality looks like a rather unnatural condition to consider: one might naturally have felt hopeful that regularity would hold, but why would one specifically feel hopeful for the scenario that if regularity fails, $\mathbb{P}_{\mathcal{C}}$ nonetheless still almost surely assigns 0 or 1 to every member of $\mathcal{C}$, even if the 0 or 1 assigned is contrary to the result of deterministic verification of whether or not the input value belongs to each member of $\mathcal{C}$?

This naturally leads to the question of whether there is some way of understanding the self-conditional triviality property that makes it more "natural" than first appears. I will present a result in this direction.

This notion of self-conditional triviality has been studied before (under different terminology from my own here), particularly in papers by Patrizia Berti and Pietro Rigo[3] where various conditions, including necessary and sufficient conditions, for self-conditional triviality are given.

My result that I will present is a further necessary and sufficient condition for self-conditional triviality.

My result

From now on, I will assume some knowledge of measure theory, but I will also give a crude heuristic description of what I am presenting.

Write $\sigma(\mathcal{C})$ for the σ-algebra generated by $\mathcal{C}$; as a crude heuristic description: if $\mathcal{C}$ is a given collection of available "test sets", then $\sigma(\mathcal{C})$ is the collection of all sets $C \subset X$ for which one, in effect, has access to the knowledge of whether a given point in X lies in C simply through the availability of being able to test whether the point belongs to each member of $\mathcal{C}$. (Since we assume that $\mathcal{C}$ consists of Borel sets, $\sigma(\mathcal{C})$ is a sub-σ-algebra of the Borel σ-algebra.)

The following is a well-known fact:[4]

Proposition. *There is a version of $\mathbb{P}_{\mathcal{C}}$ such that for each $B \subset X$, the map*

$$\begin{aligned} X &\to [0,1] \\ x &\mapsto \mathbb{P}_{\mathcal{C}}(x)\big(B\big) \end{aligned}$$

is $\sigma(\mathcal{C})$-measurable.

This "$\sigma(\mathcal{C})$-measurability" means that the set of points x for which $\mathbb{P}_{\mathcal{C}}(x)\big(B\big)$ lies in any given subinterval of $[0,1]$ is a member of $\sigma(\mathcal{C})$. A crude heuristic interpretation is that by being able to test whether a given unknown value x is in each member of $\mathcal{C}$, one has access to the knowledge of what the probability distribution $\mathbb{P}_{\mathcal{C}}(x)$ is, despite not necessarily knowing the exact value of x.

Now if we refer to $\mathbb{P}_{\mathcal{C}}$ as a "posterior probability distribution from the prior $\mathbb{P}$ given $\mathcal{C}$", then we can iterate this procedure by taking the resulting "posterior" as a new "prior":

Definition. An *iterated posterior probability distribution* (*IPPD*) *from $\mathbb{P}$ given $\mathcal{C}$* is a function

$$\mathbb{P}_{\mathcal{C},\mathcal{C}} \colon X \times X \to \{\text{probability distributions on } X\}$$

such that for $\mathbb{P}$-almost every $x \in X$, the map $y \mapsto \mathbb{P}_{\mathcal{C},\mathcal{C}}(x,y)$ is a version of $\big(\mathbb{P}_{\mathcal{C}}(x)\big)_{\mathcal{C}}$.

[3] 0–1 laws for regular conditional distributions, *Annals of Probability* **35**, 2007; A Conditional 0–1 Law for the Symmetric σ-field, *Journal of Theoretical Probability* **21**, 2008.

[4] In fact, it comes directly from the proof of the theorem that defines $\mathbb{P}_{\mathcal{C}}$.

Since, without loss of generality (according to the above Proposition), $\mathbb{P}_{\mathcal{C}}$ can be defined such that $x \mapsto \mathbb{P}_{\mathcal{C}}(x)\big(B\big)$ is $\sigma(\mathcal{C})$-measurable for each B, it seems reasonably natural to hope that one can find an IPPD $\mathbb{P}_{\mathcal{C},\mathcal{C}}$ such that $(x,y) \mapsto \mathbb{P}_{\mathcal{C},\mathcal{C}}(x,y)\big(B\big)$ (as a map from $X \times X$ to $[0,1]$) is $(\sigma(\mathcal{C}) \otimes \sigma(\mathcal{C}))$-measurable for each B; the crude heuristic meaning of this would be: Given an unknown pair of points x and y in X,

- if I can test whether x is in each member of $\mathcal{C}$,

  $\rightarrow$ which implies that I have access to the knowledge of what the probability distribution $\mathbb{P}_{\mathcal{C}}(x)$ is (for a specified $\sigma(\mathcal{C})$-measurable version of $\mathbb{P}_{\mathcal{C}}$),

- and I can also test whether y is in each member of $\mathcal{C}$,

  $\rightarrow$ which, likewise, implies that for any specified probability distribution $\mathbb{O}$ on X (and any specified $\sigma(\mathcal{C})$-measurable version of $\mathbb{O}_{\mathcal{C}}$), I have access to the knowledge of what the probability distribution $\mathbb{O}_{\mathcal{C}}(y)$ is,

then I have access to the knowledge of what the probability distribution $\mathbb{P}_{\mathcal{C},\mathcal{C}}(x,y)$ is.

Main Result (N., arXiv:1511.08864, 2017)**.** *$\mathcal{C}$ is self-conditionally trivial with respect to $\mathbb{P}$ if and only if there exists an IPPD $\mathbb{P}_{\mathcal{C},\mathcal{C}}$ from $\mathbb{P}$ given $\mathcal{C}$ such that for each $B \subset X$, the map $(x,y) \mapsto \mathbb{P}_{\mathcal{C},\mathcal{C}}(x,y)\big(B\big)$ is $(\sigma(\mathcal{C}) \otimes \sigma(\mathcal{C}))$-measurable.*